DOCUMENT RESUME

ED 191 179 EA 012 931

AUTHOR Pincus, John, Ed.; And Others
TITLE Educational Evaluation in the Public Policy Setting.
INSTITUTION Rand Corp., Santa Monica, Calif.
REPORT NO ISBN-0-8330-0215-5; Rand-R-2502-RC
PUB DATE May 80
NOTE 92p.
AVAILABLE FROM Publications Dept., The Rand Corporation, 1700 Main St., Santa Monica, CA 90406 ($7.00)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Educational Assessment; Elementary Secondary Education; Evaluation Methods; Federal Aid; *Federal Programs; *Government School Relationship; *Program Evaluation; *Public Policy; Research Design; Research Problems; *Social Science Research

ABSTRACT

The five essays that compose this report criticize, from federal, state, and local perspectives, current methods of evaluating government-sponsored education programs. A major recurring theme is that experimental design methods, most commonly used by the Department of Education, do not provide adequate information for policy-makers' needs. The essays recommend other methods that address policy-makers' immediate concerns, including such issues as resource use and distribution of funds, fidelity of implementation, and needs of target groups, in addition to the traditional focus on student outcomes. Underlying this theme is the view that the proper relationship between evaluation and policy-making is important, but ill understood, with the consequence that contemporary evaluation methods are frequently of little value in formulating education policy. Taken together, these essays offer a generally consistent set of views on the current state of program evaluation in the federal education system, and on developing new approaches to making evaluation more useful for policy. (Author/MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED191179

EDUCATIONAL EVALUATION IN THE PUBLIC POLICY SETTING

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

JOHN PINCUS (ED.), SUE E. BERRYMAN,
THOMAS K. GLENNAN, JR., PAUL T. HILL,
MILBREY WALLIN MC LAUGHLIN, MARIAN STEARNS,
DANIEL WEILER

R-2502-RC
MAY 1980

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Rand Corp
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EA 012 931

Rand

2

This research was supported by The Rand Corporation as part of its program of public service.

Library of Congress Cataloging in Publication Data
Main entry under title:

Educational evaluation in the public policy setting.

([Report] - Rand Corporation ; R-2502-RC)
Bibliography: p.
1. Education and state--United States. 2. Education--United States--Evaluation. I. Pincus, John A.
II. Series: Rand Corporation. Rand report ;
R-2502-RC.
AS36.R3 R-2502 [LC89] 081s [379.1'54] 80-13656
ISBN 0-8330-0215-5

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

PREFACE

The five essays in this report have the common goal of improving current methods of evaluating government-sponsored education programs. Four of the essays, Chaps. 2 through 5, were originally written for separate purposes and audiences, but deal with closely related themes. The Rand Corporation supported the collection of those essays into a single volume, and commissioned an introductory essay (Chap. 1, by John Pincus) that summarizes the findings and general tenor of the other chapters and urges a broader concept of evaluation than that currently prevailing.

Chapter 2, by Sue E. Berryman and T. K. Glennan, Jr., is a product of work performed by Rand's Educational Policy Research Center for the Office of the Assistant Secretary for Education, U.S. Department of Health, Education, and Welfare. It reviews and recommends improvements in U.S. Office of Education program evaluations.

Chapter 3, by Milbrey Wallin McLaughlin, was originally a paper delivered to an annual meeting of the American Political Science Association. It is based on a Rand study, conducted for the Office of Education, of federally funded innovative projects in some 200 school districts over a four-year period.

Chapter 4, by Paul Hill, was also written for the Educational Policy Research Center. It distills the experience of directing a major study of Title I of the Elementary and Secondary Education Act, commissioned by and for Congress.

Chapter 5, by Daniel Weiler and Marian S. Stearns, recommends changes in California's state-sponsored evaluation policy. It was written in the authors' capacity as members of the California Commission on Education Management and Evaluation, for the use of the California State Board of Education and other state policymakers, including the executive and legislative branches.

Although Chaps. 2 through 5 were written at different times and with little or no consultation among their authors, the reader will perceive their complementarity. Taken as a group, they present a generally consistent argument for changes in the theory and practice of program evaluation at the federal, state, and local levels. The reader will also note occasional divergences of opinion regarding the proper future direction and emphasis that evaluation should take. Chapters 1 and 3, for example, advance viewpoints about current evaluation practices that emphasize issues not discussed in the other essays. Such divergences are inevitable in a collection of separate pieces that deal with both short-run and long-run phenomena and encompass different fields of vision. The essays differ in the perspectives taken by their authors, then, but are not inconsistent in their basic findings.

SUMMARY

The essays that compose this report criticize, from federal, state, and local perspectives, current methods of evaluating government-sponsored education programs. A major recurring theme is that experimental design methods, most commonly used by the U.S. Department of Education, do not provide adequate information for policymakers' needs. The essays recommend other methods that address policymakers' immediate concerns, including such issues as resource-use and distribution of funds, fidelity of implementation, and needs of target groups, in addition to the traditional focus on student outcomes.

Underlying this theme is the view that the proper relationship between evaluation and policymaking is important, but ill understood, with the consequence that contemporary evaluation methods are frequently of little value in formulating education policy.

Chapter 1 of the report summarizes the main findings of Chaps. 2 through 5, and discusses a possible new approach to evaluation through the social effects of government programs.

Chapter 2 is a critique of the experimental design approach, with recommendations for new perspectives on evaluation.

Chapter 3 analyzes federal program evaluation from the local perspective, and argues that present evaluation methods are generally not useful from the local perspective. It also points out the absence of any evaluation approach that can successfully deal with the sources of local variations.

Chapter 4 describes an Executive Branch attempt to evaluate Title I of the Elementary and Secondary Education Act on behalf of Congress. The strategy adopted, which was generally successful, was to define the aims of the evaluation in light of Congress's policymaking authority and concerns.

Chapter 5 discusses, for an audience of state government legislators and administrators, how education programs can be more successfully evaluated at the state level.

Taken together, these essays offer a generally consistent set of views on the current state of program evaluation in the federal education system, and on developing new approaches to making evaluation more useful for policy.

ACKNOWLEDGMENTS

The authors are grateful to The Rand Corporation for underwriting the costs of publishing this report. Berryman, Glennan, and Hill acknowledge the support of the Office of the Assistant Secretary for Education, HEW, for the original preparation of Chaps. 2 and 4, respectively. Weiler and Stearns acknowledge the collaboration of members and staff of the California Educational Management and Evaluation Commission in preparing Chap. 5.

All the authors express their appreciation to Professor Richard F. Elmore of the University of Washington, and to Paul Berman, a Rand colleague, who reviewed the manuscript and made valuable suggestions for its improvement.

The authors of each chapter have acknowledged individual contributions in the beginning of their chapters.

CONTENTS

PREFACE iii
SUMMARY v
ACKNOWLEDGMENTS vii

Chapter
1. THE STATE OF EDUCATIONAL EVALUATION: REFLECTIONS AND A SUMMARY 1

2. AN IMPROVED STRATEGY FOR EVALUATING FEDERAL PROGRAMS IN EDUCATION 11
    Introduction 11
    The Setting for Federal Education Program Evaluations 13
    Dominant Evaluation Strategy 17
    An Alternative Strategy for Evaluating Education Programs 24
    Implementing the Evaluation Strategy 34
    Bibliography 37

3. EVALUATION AND ALCHEMY 41

4. EVALUATING EDUCATION PROGRAMS FOR FEDERAL POLICYMAKERS: LESSONS FROM THE NIE COMPENSATORY EDUCATION STUDY 48
    Introduction 48
    The NIE Study and the Problems It Faced 49
    The Problems and How NIE Dealt with Them 51
    Summary and Conclusion 73

5. THE USES AND LIMITS OF EDUCATION EVALUATION AT THE STATE LEVEL 77
    Foreword 77
    Conflicting Needs and Perspectives 78
    Appropriate Expectations for Evaluation 79
    Improving the Usefulness of Evaluations 82
    Evaluation Policies and Priorities 84

Chapter 1

THE STATE OF EDUCATIONAL EVALUATION: REFLECTIONS AND A SUMMARY

by

John Pincus

It often happens in the development of scientific methods that a number of people point out contemporaneously the defects of some existing approach to research. These researchers, or others, may then develop improved methods for the same kinds of research tasks. This volume brings together the work of six researchers who point out, from different perspectives, the weaknesses of the experimental model for evaluating social programs in general, and education programs in particular. But this, of itself, is not an original contribution. Other writers, as the bibliography to Chap. 2 indicates, have ably discussed the deficiencies of the experimental design model for use in social program evaluation.

It is not in that quarter, therefore, that these essays lay their claim to novelty, but in their focus on the relation between educational evaluation techniques and policymaking, throughout the federal system. Government agencies that commission program evaluations have certain views about the role of policy. These views may affect their approach to evaluation, which will in turn affect policymakers' perceptions of what did happen and of their own ability to influence what will happen in government-funded programs. In other words, evaluators have their views about appropriate evaluation methods, and these, when applied, will affect the usefulness of program evaluation for public policy.

Therefore, the nature of the relation between policymaking in general and evaluation policy in particular serves to define the perspectives of both the evaluator and the policymaker.

A central theme of these essays is that evaluation policy and findings are often affected by researchers' tendency to apply the experimental design paradigm to most education program evaluations, thereby creating persistent tension between researcher and policymaker. In the later 1960s, shortly after the federal government's large-scale entry into subsidy of education programs, the Office of Education began to call on researchers who were schooled in experimental design to evaluate these programs—notably Title I of the Elementary and Secondary Education Act (ESEA)—that were designed to improve the schooling and economic opportunities of children from poor families. The researchers tended to strive for appropriate quasi-experimental conditions, while the policymakers sought something more useful than the researchers' usually inconclusive findings about test scores.

Most policymakers want their programs to succeed; but most "scientific" evaluations address effects and indicate that student outcomes as measured by test scores, dropout rates, and other such measures appear to be little affected by new government education programs. Such reports of "no significant effect" are generally unaccompanied by useful recommendations for program improvement or policy change. Meanwhile, policymakers seek to know not only about effects, but also about what is going on in the program: how the resources are being used, whether implementation corresponds to program intent, and who is benefiting from program resource use. In effect, what can result is a "dialog of the deaf," in which neither party understands the other's premises. Is it possible to reduce these tensions and improve the utility of evaluation to public policy? This is the main theme that recurs in each of the next four chapters.

Berryman and Glennan, in their essay "An Improved Strategy for Evaluating Policy in Education," argue that appropriate evaluation methods for a federal program cannot be defined without reference to the policymaking process, which is political in that it involves real value conflicts. Usually, because no one party can impose a solution of the conflicts, policymaking becomes a process of compromise—a mechanism, in the authors' words, for resolving differences. If this is true, then evaluations are likely to be most useful if they address different outcomes and processes, each responding to some party's interests in the policymaking process.

The major problem with educational evaluation that Berryman and Glennan cite is its reliance on the experimental model as a strategy for evaluating the large federal education programs. This model is most suitable for estimating program effects for small project grant programs, or for estimating the feasibility of obtaining treatment effects under optimal conditions, such as controlled conditions.

Policymakers are the major users of program information. Their information needs should determine what evaluation studies should be conducted. The experimental model is too limited to fit those needs. For the large federal education programs, in addition to knowledge of effects, policymakers need a broad range of *descriptive* information, *interpretive* information about program effects, and information about mature as well as new programs.

During its life cycle, a program's implementation and effects change and policymakers' information-needs change accordingly. Therefore, policymakers need a multiyear information plan that reflects their information needs—at any one time, and as they change over time.

Berryman and Glennan base their recommendations for a new approach to program evaluation on the evaluation plan for the newly enacted Education for All Handicapped Children Act and on the NIE compensatory education study (see Chap. 4 below). Their criteria for improving program evaluation include:

- Recognizing the primacy of policymakers' needs.
- Assessing the reasonableness of program assumptions.
- Interpreting how elements of the program combine to produce program outcomes.
- Documenting the extent of diversity in program services, thereby demonstrating how a program actually works.
- Focusing on program description as well as program effects under optimal and typical conditions.
- Timing studies to accord with the life cycle of the program. This in turn implies a long-term plan of evaluation studies, a plan that is negotiated annually among the policymakers, the major users of the study results.
- Conducting a large number of small studies with limited objectives, instead of fewer studies with multiple purposes.

Berryman and Glennan recommend that this new strategy be carried out by a centralized staff for evaluating federal education programs. The staff would report to the Secretary of Education, and report study results simultaneously to Congress and the Secretary.

In Chap. 3, Milbrey McLaughlin addresses the same theme—the weaknesses of federal evaluation policy—but from a different slant. She is not explicit about the roles of researchers and clients, and leaves the impression that better understanding of phenomena is the aim. Implicitly, her client is a single agency that makes grants for advancing educational research. In McLaughlin's perspective, local factors dominate the success or failure of federally funded projects. These factors are so complex, and often so idiosyncratic, that it is usually impossible to evaluate a program independently of the local settings; and she argues that evaluation in the local setting requires development of new approaches both as to theory and measurement. Her critique of existing approaches extends beyond the experimental model to include the economist's production function approach and the use of such statistical methods as correlation and regression. On the basis of several years of research into the local performance of federally funded projects, she argues that current evaluation methods cannot deal effectively with qualitative elements that, in part, define the local setting and the nature of local participation.

McLaughlin believes that knowledge of the local setting and of implementation processes is valuable because it undermines the case for existing approaches to evaluation, and challenges investigators to develop new ones and to use evaluation for purposes now unheeded, such as defining local settings that are conducive to reform.

From a certain perspective, McLaughlin's argument asks for more than is contained in the alternatives proposed by Berryman and Glennan, because their essay does not attempt to solve simultaneously the fact of local variation and the federal need for general findings as to outcomes. But, as McLaughlin points out, the federal government needs to know how money was spent, whether programs were implemented, and what the outcomes were thereafter. This kind of information, however imperfectly it may render reality, is a political necessity, because it gives policymakers the closest approximation now available in the attempt to understand what is happening. This approximation may be, as McLaughlin implies, often remote because the paradigms are wrong. Yet we know from experience in many fields that old paradigms rarely wither away. They must first be overthrown by new ones—a condition that in evaluation, as McLaughlin points out, seems far away.

In Chap. 4, "Evaluating Education Programs for Federal Policymakers," Hill describes how research decisions were made for the one education evaluation that Congress has found most useful, the NIE Compensatory Education Study. In responding to a Congressional mandate to evaluate a national program, researchers must overcome five basic problems: translating the often tangled language of the mandate into a set of researchable questions; ensuring that the research makes a fair summary judgment of the program under evaluation; overcoming Congressional distrust of researchers; managing advocacy pressure from interest groups and individual members of Congress; and producing reports that Congressional staff can use.

Hill says that Congress, not researchers, ultimately evaluates national programs. He argues that the researcher's job is to provide a background of commonly accepted basic facts, and predictions of the results of available options, to facilitate the political decisionmaking process. Because Congress is normally divided about a program's objectives, it is seldom possible to use a single criterion for program effectiveness. In the case of Title I, some members of Congress were principally concerned about the funding patterns it created; others wanted to upgrade the quality of instruction provided to disadvantaged children; and still others cared only about raising disadvantaged children's academic achievement. Sole reliance on the last criterion would deprive many members of Congress of the information they needed to evaluate the program.

Hill argues that research planning must start with a careful assessment of Congress's information needs. In the case of Title I, members of Congress knew very little about how the program operated, what students were served, and what services were delivered. Providing simple descriptive information thus became one of NIE's key objectives. Consultations with Congressional staff also helped NIE identify the issues that were most likely to arise in Congressional debate. Advocacy groups presented a far broader agenda of issues for the study to consider, but NIE concentrated its resources on illuminating those issues most likely to be important in Congress. That strategy provoked criticism and opposition from some interest groups; the study staff survived the resulting bureaucratic pressure by relying on Congressional support.

Major evaluations are subject to political pressures from all sides, and if the researchers capitulate to those pressures, they lose all claim to objectivity. Research results will be discredited if they fail to meet normal professional standards of validity and fairness. At the same time, the political pressures often convey information about what issues are important and how research results will ultimately be used. Hill argues that researchers must understand their political environment, not ignore it, and take care to avoid presenting results in ways that are too easily adapted to propaganda purposes.

In adopting these perspectives, researchers sacrifice much of their independence in identifying research problems, and must exercise uncommon restraint about expressing their own opinions. As Hill notes, supporting the Congressional decisionmaking process is a conservative activity that leaves no room for the advocacy of fundamental policy changes that Congress would not or could not consider. Though the researcher can avoid being captured by the contending Congressional and interest-group factions, he is, ultimately, supporting the existing regime.

In many ways, Hill's experience was unusual; executive branch researchers rarely work for Congress unfettered by their agency's control, as Hill did with respect to his employer, NIE/HEW. Nonetheless, his experience included many of the elements common in any large-scale government evaluation: competition for a voice in research design and for control of information flows; the incentive to satisfy client needs; the need to disarm or at least consult with potential adversaries (who can include virtually any group not consulted with); the desire to satisfy research-staff aspirations for quality and pertinence in the evaluation; and our familiar theme of finding the appropriate research design.

Chapter 5, "The Uses and Limitations of Education Evaluation at the State Level," takes up many of the issues raised for the federal level in Chap. 4, notably the uneasy relation between policymakers' needs and evaluators' working style. It also discusses the issues of implementation and evaluability discussed in Chap. 3. The main aim of the original paper, prepared at the request of the California State Board of Education, is to encourage more effective working relations between state government and education program evaluators. The authors lay most stress on the frequent inability of evaluators to meet policymakers' needs for information of a type that will help their deliberations. Weiler and Stearns suggest a lessening of emphasis on large summative studies and on experimental research design. They also urge more collaboration among evaluators and government agencies to establish agreement about goals of the evaluation and appropriate research designs. In effect, they ask the evaluator not to promise too much, and the state government not to ask or expect too much from evaluations. Program evaluation, according to Weiler and Stearns, has only a modest record of achievement. More is to be gained, they believe, from identifying problems that might require state or school district action, or whose identification will support management decisionmaking, than from attempting to evaluate the effectiveness of program practices or cost-effectiveness of the programs as a whole.

Like Hill, the authors focus primarily on evaluation as a tool for government. Like him, they essentially stress simplicity of research design, avoidance of elaborate evaluation of cost-effectiveness, simplicity of language in reporting results, and clear communications with the client, all intended to create a useful atmosphere for policymaking.

Collectively, these four essays find fault with current evaluation methods, each from a different perspective, and call for improvements. Berryman and Glennan deplore the experimental model as a paradigm for evaluating federal education programs, and suggest other approaches to research that will aid policymakers. McLaughlin stresses the preponderant importance of idiosyncratic local factors in success or failure, and calls for new paradigms of evaluation that take idiosyncratic and qualitative factors into account. Hill and Weiler and Stearns stress the relation between the evaluators and the policymaking client. They call for clarity, simplicity, and modesty in research design and reporting.

Together, these four essays constitute a strong attack on the use of experimental research design in inappropriate situations, such as large-scale educational interventions. They also stress the importance of designing evaluations so as to meet policymakers' needs, and find the resolution of these issues in what is essentially an unassuming response. They say in effect that social science should pull in its horns, and should forget about global assessments of program effectiveness. Instead, they aver, social science researchers should find out what policymakers need to know: descriptions of the program and of resource use, evidence on implementation, some interpretation of program effects, including student outcomes, and other variables. Finally, they emphasize that a number of modest studies of different phenomena of interest are preferable to a single study aimed at one or two central issues, programwide.

Collectively, they call for a retreat from the somewhat overambitious pretensions of social science in the earlier years of evaluation studies, and a proposal that evaluation be considered primarily an information source for managers and policymakers, describing and interpreting different aspects of the program.

The approaches that they counsel are based on current awareness of the atmosphere in which large-scale educational interventions take place. In this atmosphere, the nature of the treatment provided is often unknown, and is rarely the same from site to site. Therefore, outcomes, even when measurable, can rarely be associated with treatment.

To the extent that the essays mention specific outcomes, they focus on service delivery and student achievement. Berryman and Glennan point out that it is often uncertain at the start of a program whether the assumptions that underlie program goals are reasonable; one "objective" may be to find out whether program goals can be met at all. In a way, their approach is a halfway house down the road of social program evaluation, stopping short of broader goals for evaluation.

Social science evaluation of federal and state programs is in one respect a curious blend. Immodest in its claims, as the authors of this volume document, it remains strangely modest in its aspirations. And, by omission at least, these essays share that modesty. Title I, for example, was not primarily designed, as Hill points out, to give school districts money, to provide services to the poor, and to promote their development. These were intermediate goals. The prime goal of Title I was to help advance the social and economic status of the poor in America, largely the minority poor. Yet of the tens of millions that have been spent on Title I evaluation, no money has been devoted to finding out how well that goal has been met. Children who were ten years old when the program began are now twenty-four, yet researchers are still treating test scores as outcome measures.

This practice of selecting intermediate goals, such as those cited by Hill, understandably dominates in the early years of a program. During the period when the dominant goals are yet to be realized or even tested, the attention of public servants is concentrated on whether funds have been properly spent, on how programs are carried out, and how children seem to be affected in ways that can be measured. Later on, a time comes when the children who have been exposed to Title I, or other programs, leave school, and go on to college, work, or unemployment. These are the arenas where the worth of the program is now tested every day, and where the public service should now be commissioning evaluations if it seeks to fathom the effects of the program.

The main difficulty of evaluations, from a certain perspective, is that they fail to examine the social consequences of social programs. What good does it do, in the long run, to know about resource-use, fidelity of implementation, or test scores? People who are poor, receive no extra resources, and have average test scores often perform well in society, while others with the same credentials perform poorly. The issue for society is the social and economic consequences of educational reforms. Measures such as resource-use and test scores "explain," in the statistical sense, little about social and economic futures, once we take into account family background of the student. The remaining variations in test scores explain little about career patterns.

There are other measures of the social effects of compensatory education programs besides income and measures of social status. For example, how do large interventions affect the behavior of school systems—of teachers, administrators, school boards, students, and parents—and how is the relation between schools and society consequently affected? How do people judge success and failure in such programs: by (1) perceived increases or reductions in school discipline problems (many Americans consider poor discipline to be the dominant problem of the public schools); (2) perceived changes in academic performance; (3) increases in teacher willingness to work in these programs; (4) expression of children's satisfaction or dissatisfaction with programs? It is arguable that popular perceptions should not be considered a social consequence of a program. On the other hand, pervasive perceptions may be viewed as intermediate outcomes, having their effect on the atmosphere in which programs are judged and their graduates dealt with, as well as affecting the atmosphere in which the program's future is debated. Other measures are possible, but will not be discussed here.

The points that emerge from these considerations are simple. Education is a social activity. The deliberate reform of education has a social purpose, and may also have other purposes—administrative, pedagogic, financial, or political. These aims in turn, if successful, have effects on the social system. Therefore, to evaluate an educational program well is not possible in the long run unless one measures its effects on society, in light of the aims that different parties foresaw in launching the program. Cases in point, in addition to Title I, include:

- *The Education for All Handicapped Children Act.* Its aims are presumably to convince parents of handicapped children that society is willing to accommodate their children's needs and to increase their capacity to function effectively in society.
- *Headstart.* Its original proximate aim was to demonstrate that young children from poor families could handle schooling better if they started school at age 3 or 4. The long-run aim was the same as that of Title I. When the data did not confirm that the proximate goal was being attained, the program was continued anyway to demonstrate the federal government's concern for poor minority children and their parents, under the general rubric of increasing "social competence." The long-run goal was presumably unchanged, except that "social competence" instead of academic ability now became the catalyst.
- *Bilingual Education.* The program began by financing the extra costs of bilingual classes, so that Spanish-speaking students could use Spanish as a transition toward learning English. Its aims were to demonstrate federal interest in Spanish-speaking people, and give them a greater sense of ownership in local school systems. A subsidiary aim was to advance the children's opportunities in life. Later, Spanish-speaking interest groups added a new aim: the expression of the political independence and authority over the program of organized people of Spanish-speaking background.

There are other programs—Vocational Education, compulsory schooling to age 16 or 18, the development of state colleges and universities—whose aims have evolved considerably, and illuminate the ways in which program aims shift in response to both early experiences with the program and to changing social forces.

The field of vision, in such an approach, is the dominant factor in determining the success of a social program. What is required is an unusual mix of research talents: the historian, to determine the original aims and the changes therefrom; the sociologist, to identify and measure how such a program might affect social systems; the economist, to measure effects on earnings and on costs to society; the political scientist, to identify the dynamics of changing aims; the organization theorist, to identify the nature of institutions' effects on meeting and forestalling the aims.

Needless to say, this approach is not limited to educational initiatives. Such an evaluation approach could cast light on the Agricultural Adjustment Act, founded in 1933 to save the family farm. From the perspective of nearly half a century, one can see that the program has primarily subsidized large-scale commercial farming, and presumably thereby contributed to the demise of the family farm. Similarly, federal urban policies have in many cases contributed to the economic decline of cities. The study of such programs as these might clarify the defects of the founding assumptions (as Berryman and Glennan suggest), and the nature of the administrative and political processes that led the programs in the direction they achieved.

Why should such evaluations be pursued, and for whom and how should they be done? The case for doing such evaluations relates to the ultimate argument for evaluation: to demonstrate the relationship between intervention and outcome. Most evaluations currently settle for proximate goals, which are necessary but generally bear unknown relations to ultimate goals.

The audiences are not primarily program administrators or legislators, although the latter may eventually benefit to the extent that evaluations cast light on systematic sources of difficulty in meeting goals. The primary audiences are the administrators who set national policy from the White House or from Cabinet positions, since they are responsible for formulating national policy; and the public at large, because the policies affect them and are financed by them.

There is no general answer to how such studies should be done. If we are dealing with established programs (agricultural adjustment, vocational education), historical, social, and economic research should dominate, because the data each discipline needs to define changing goals and to measure changing economic and social phenomena are available. If the program is relatively new (bilingual education, for example), then the approaches suggested by the other essays in this volume become appropriate, with the addition of data collection on program origins and of a research design for the longer-term social evaluations.

This approach makes an appraisal of social effects, in light of changing program goals, the prime long-term aim of evaluation, and appraisal of proximate effects the dominant task of the early years of a program. During the early years, the goal is essentially technical and political: Is the job being done the way it is supposed to be, both as to flow of funds and direction of human energies? What relation does it show to proximate measures of outcome? Does the evaluation provide the policymaker with the information needed for making decisions within his or her sphere of action?

This approach is clearly more ambitious than current evaluation policies envisage. It is also difficult, because time spans are long and because, in dealing with social causality, it is difficult to separate the effects of different influences on a given outcome, particularly if certain influences are difficult to measure or even to define (for example, the combined effect of judicial decisions on minorities' economic and social status). Yet such evaluations have been done with careful conceptual approaches and effective use of existing data bases (for example, the studies of Freeman, Smith, and Welch),[1] for particular questions of social interest, such as the effect of schooling on earnings.

---

[1] Richard Freeman, *The Overeducated American*, Academic Press, New York, 1976; James P. Smith and Finis Welch, *The Overeducated American? A Review Article*, The Rand Corporation, November 1978.

Because this approach is both long-term and ambitious, it may be argued that the activity is really broad social research rather than evaluation. In part such a judgment would depend on who commissioned the research and for what purpose. If it were commissioned as a long-term evaluation of Title I, the judgment might differ from that of an independent academic study of social trends in education. In any case there would always be some evaluative component, and probably some broader research aspect.[2]

Up to now, there has been relatively little demand for such an approach to evaluation, largely because those whose task it is to pursue the proximate goals also commission the evaluations. In such a setting, they must require evaluation to focus on those same objectives.

One difficulty, which Hill's essay points out, reflects another reason for the present situation with evaluation. Often, the program authorization mentions the social goal only in a preamble or not at all, so that a serious evaluator must, given the current situation, confine the evaluation to stated purposes. If the implicit and the stated purposes differ, a conscientious evaluator cannot pursue the implicit aims without a mandate to do so. It should be noticed that three of the other essays in this volume suggest that aims or assumptions be reviewed with the client. Such an approach would allow the introduction of social effects as a subject of evaluation. However, in the types of situations that these authors describe, it is late in the game to suggest that a client reorient so profoundly the aims of the investigation.

This reorientation, which encompasses the reforms discussed in Chaps. 2, 4, and 5 and aims at some of the further reforms called for in Chap. 3, is no small task. Many programs may prosper only because their social effects are never tested, or at least never accepted. Others that are widely criticized may in fact have largely favorable social effects. There are vested political, social, and economic interests behind the reigning attitudes toward programs, and even proximate objectives can take on their own independent identities and their own supporters. The perspective of policymakers is constrained by their terms in office; therefore, social-effects studies can perhaps be undertaken only for established programs. Finally, the foregoing issues aside, policymakers may prefer, with some reason, their own judgment of social effects to those of evaluators.

But even if all these objections and others carry weight, it still remains true that the main methodological objection to present-day evaluation of social programs is that, with the exception of some that have direct economic aims (such as job programs), they are not evaluating what the programs are ultimately trying to achieve. It is also true that, ultimately, success or failure in achieving these aims is the test that such programs must pass.

To the extent that evaluation turns to such aims as these, the prescriptions set forth in the accompanying essays are all the more cogent. The present essay simply lends further insight to the emphasis on mutual understanding of goals and the assumptions that underlie them; on abjuring experimental design, in most cases, except where government consciously mounts experimental efforts; on reporting results clearly and simply; and on breaking large topics up into a number of smaller related research initiatives.

---

[2] See also Scarvia S. Anderson and Samuel Ball, *The Profession and Practice of Program Evaluation*, Jossey-Bass, San Francisco, 1978, pp. 9-11.

Finally, it is inappropriate to judge certain programs by their social effects, because their aims are not primarily social. A case in point is Title IVC of the Elementary and Secondary Education Act, whose proximate aim—encouraging school districts to try out innovative educational methods—is probably the dominant one. Its social aim would probably be defined by its effects on teachers and school systems, effects having to do with receptivity to change and willingness to introduce new methods as appropriate. Even less directly, there would be similar effects on children and on society. But these effects would be difficult to measure and presumably modest in size compared with other effects. In such cases as this, the proximate goals are probably the only ones that can be measured.

Consequently, this author anticipates that efforts to include a judgment of social effects in program evaluation would result, not in abandonment of proximate goals, but in a spectrum of evaluation types depending on the age and nature of the program. In some cases there would be no attempt to measure social effects, in others no effort to measure proximate ones, and in still others both proximate aims and longer-term social effects would be objects of inquiry.

This is no counsel of perfection, designed to nullify efforts to improve other defects of contemporary evaluation. The improvement of methods for evaluating proximate goals will remain a prime task for evaluation strategy since, under the system loosely sketched here, each social program would be evaluated on the basis of proximate goals, particularly in its early years (and in many cases indefinitely). And many programs, because of their complexity, absence of data, or frequent changes in goal, will remain elusive to the evaluator of social processes. In that case, proximate criteria must substitute for underlying aims.

For these reasons, and because the presentation of a more useful way to conduct research must, to borrow the words of Joseph Conrad, carry its own justification in every line, the following essays deserve the reader's attention. Collectively, they endeavor to show us a better way of assessing the educational world we are trying to create; if we are unable to see that world clearly, we can make only groping efforts to shape it to our needs.

Chapter 2

AN IMPROVED STRATEGY FOR EVALUATING FEDERAL PROGRAMS IN EDUCATION

by

Sue E. Berryman

Thomas K. Glennan, Jr.

INTRODUCTION[1]

Evaluations of federal education programs began soon after the passage of the Elementary and Secondary Education Act (ESEA) in 1965. Systematic evaluations of federal education programs date from 1970, with the establishment of the Office of Planning, Budget and Evaluation (OPBE) in the Office of Education (OE). (The OPBE recently changed its name to the Office of Evaluation and Dissemination.)

Although evaluation of educational programs has become an established federal activity, a series of recent events has stimulated us to review the strategy that has governed evaluation of these programs.

- Participants at a recent Rand conference on federal aid to education were struck by two contrasting observations.[2] One was that after ten years the OE was beginning to find that states and school districts were in reasonable compliance with Title I of ESEA. The other was that, from the beginning, the evaluations of Title I have focused on children's academic achievement. The participants questioned whether it was reasonable to expect clear educational results from a program early in its development. The federal government's strategy for evaluating Title I had seemed to reflect little appreciation of the implementation stages that federal education programs go through. Was federal program evaluation in education generally as unaware of the implementation processs as it seemed to be in the Title I case?
- Over the last four years Congress has mandated two evaluations—one of compensatory education, the other of vocational education—that have some unusual features. First, the mandates make it clear that Congress is the client for these evaluations. For example, in these two cases the manag-

---

[1] Michael Timpane participated actively in this research until he left Rand for the National Institute of Education. For their comments and suggestions, we also thank Mark Abramson, Joel Berke, Rudy Cordova, Lois-ellin Datta, Richard Elmore, John Evans, Jerry Fletcher, Edward Glassman, Jack Jennings, Abdul Khan, Richard Light, William Lobosco, Garry McDaniels, Constantine Menges, Fritz Mulhauser, Michael O'Keefe, Marshall Smith, Steve Weiner, Joseph Wholey, and Carl Wisler. We owe special thanks to Paul Hill of Rand, who offered detailed and constructive criticism of earlier drafts.

[2] The conference papers were published in book form. See Timpane (1978).

ing agency for the evaluation is required to go to the Congress without prior executive branch review. Second, the National Institute of Education (NIE), not OPBE (the agency traditionally responsible for evaluating OE programs), is responsible for managing the evaluations. Third, the mandates request a broad range of information. In addition to the usual information about achievement outcomes, the mandates request analysis of alternative ways of distributing federal funds and identifying children in need. Fourth, the deadlines for study results are established to accord with the reauthorization of the individual programs.[3]

- In 1977, the General Accounting Office (GAO) issued a report on problems with the evaluation of OE programs.[4] Congressional staff members told the GAO researchers that OE evaluations have little impact on legislation. They complained that OE studies are not timed to coincide with the legislative cycle; OE efforts to interpret data and highlight important findings are insufficient; and OE briefings for Congressional committee staff are not frequent enough.
- Leading evaluation researchers have increasingly questioned evaluation practices and the norms that have governed them. For example, Campbell (1974), who is strongly identified with the experimental approach to evaluation, has argued for systematic, qualitative information to supplement data yielded by formal experimental designs. He recommends obtaining information on a program's content, its implementation, its community context, and its clients. Cronbach (1974), citing evidence on the sensitivity of human behavior to variations in context, argues for "interpretations in context" instead of generalizations across contexts. In their assessment of the Follow Through evaluation, House et al. (1978) argued for several changes in evaluation practice. For example, they argued for evaluations that are sensitive to local conditions, and against massive evaluations with narrow outcome measures.
- Educational policymakers and evaluators, including critics, generally agree that OPBE, the Office of Education's major evaluation unit, has used methodological practices espoused by a substantial number of scholarly evaluators and has protected the process from political bias. The director of OPBE is a leading and thoughtful spokesman on the problems of evaluating social programs, and the staff is regarded as well qualified.

Consequently, few if any unsatisfactory evaluations of federal education programs can be attributed to incompetence or bias of those who fund or conduct them. This essay seeks to identify some of the sources of evaluation problems and to propose solutions. The research community now has over a dozen years of experience with evaluating education programs and with policymakers' responses to these evaluations. This accumulated experience can shed light on the current discontent with evaluation. The creation of a Department of Education makes this an opportune time to devise new strategies for assessing federal education programs.

---

[3] Recently (1978), HEW reacted to Congress's new research style. It persuaded Congress, in mandating a three-year study of school finance policies, to vest authority for the study in the Secretary of HEW, rather than its constituent agency, NIE.

[4] Comptroller General of the United States, *Problems and Needed Improvements in Evaluating Office of Education Programs*, General Accounting Office, Washington, D.C., September 8, 1977.

This report applies only to federal preschool, elementary, and secondary education programs; it excludes postsecondary and adult education programs, which have quite different objectives and institutional settings. Furthermore, we consider only federal information needs. State and local governments have analogous information needs and may find the report relevant, but federal and nonfederal powers and responsibilities differ, and their evaluation needs require a separate analysis.

We used a variety of sources for this essay: talks with people in the executive branch, on some Congressional staffs, and in the academic community; the evaluation literature, OPBE Requests for Proposals, OE evaluation reports, and public documents about OE evaluations; analyses of how the NIE and Congressional staffs managed the compensatory education evaluations for the Congress; and talks with research staff at the Bureau for the Education of the Handicapped as they developed an evaluation plan for the new Education for All Handicapped Children Act.

In the remainder of this essay we argue that:

- The nature of policymaking processes and federal education programs constrain what we can know and want to know about federal education programs;
- Practices associated with experiments in psychology dominate education program evaluation; these practices are not entirely appropriate for evaluating large federal education programs; and evaluations based on them do not provide sufficient information for policy decisions;
- An alternative strategy that we propose should satisfy policymakers' information needs better, a strategy consisting of: a substantially broader range of studies than those conventionally associated with evaluation: a multiyear plan of studies for the large federal programs; the timing of particular studies to accord with the programs' life cycle; and annual negotiation of the substance of the plan among major users of the evaluation, including the Congress; and
- This alternative strategy will require organizational changes in the federal education evaluation system.

THE SETTING FOR FEDERAL EDUCATION PROGRAM EVALUATIONS

Federal agencies generally use evaluation to assess the consequences of policy actions. Characteristics of both the policy process and federal education programs determine what information policymakers can use and can get. In this section we characterize the policy process and education programs and suggest their implications for what we need to know and can reasonably expect to know. We then sketch what this analysis implies for federal educational program evaluation.

Characteristics of the Policy Process and Education Programs

Our analysis rests on several assumptions about the nature of the policy process and federal education programs. We introduce these assumptions (in some cases, factual statements) briefly. They come from the authors' personal observations of public policy and from the writings of others with policy and evaluation experience (e.g., Williams, 1972; Lynn, 1973 and 1977; Morrill and Francis, 1976; Williams and Elmore, 1976; Menges, 1978), and should command general consensus among policy analysts and policymakers. We do not try to justify the selection of these particular statements, except by the test of their usefulnesss to our later argument.

The Policy Process. Seven characteristics of the policy process strongly shape the needs for knowledge produced by evaluation.

1. Policymaking is not a discrete event, but a process extending through time.

2. This process entails resolving real value conflicts. A policy action normally affects different parties differently. Since the executive branch, legislature, judiciary, and private interest groups share the power over policy outcomes (none of them has complete power), the policy process becomes a mechanism for resolving differences among them.

3. Social phenomena, such as poor children's achievement scores on reading and math tests, come to be defined as social problems through the policy process. Since the policy process involves groups with different preferences, a social problem typically takes on many definitions.

4. Social problems tend to persist and are often redefined as time passes. Problem definitions can change in response to a variety of events. There can be a change in the empirical character of the social problem. For example, if the economy has a limited number of jobs suited to teenage skills and preferences, unemployment among teenagers can fall as their numbers decrease. New knowledge can be gained about a problem. For example, it may be found that delinquency precedes and encourages dropping out of school, instead of dropping out causing delinquency. Or social priorities can change, often as the result of changes in the distribution of power and therefore in the prevailing values used to define the problem. For example, poverty was defined less as a structural problem and more as a personal problem of the poor when the Nixon administration replaced the Johnson administration.

5. Although the policy process operates to resolve fundamental value differences among the parties, empirical information affects policy outcomes. In responding to a social problem, policymakers at first know little about the problem or the likely consequence of their actions. A policy action therefore usually starts a learning process, rather than culminating it. However, since social problems tend to persist, the policy process can and tends to incorporate the public learning that it sets in motion.

6. Periodic events in the policy process force reassessment of policies. For example, Congress holds reauthorization and oversight hearings. The executive branch conducts the annual budget process, and administrations change periodically.

7. The powers and responsibilities of these political institutions determine what kinds of knowledge are needed for such reassessments. The nature of Congressional and executive responsibilities requires information about program cost, implementation, and outcomes. Congress, as lawmaker and fund allocator, can start, maintain, or change policies and programs. For any program, it can specify intent, administrative arrangements, and funding levels and formulas. The executive branch administers Congressionally mandated programs. It has the power to

write regulations that specify how programs shall be implemented. Congress oversees what the executive branch does, and can review executive regulations.

Federal Education Programs. Some characteristics listed here are attributes of federal programs in general, not merely of federal education programs. Others are found only in federal education programs.

1. Because creating federal program legislation usually requires a coalition of interests, it typically leads to programs with ambiguous and multiple objectives.
2. Because the Congress rarely ends programs, they tend to persist.
3. Both the Congress and the executive branch, however, tend to modify programs through time. The programs' funding, objectives, and administrative regulations can all change. Since policymakers tend to respond to bad news about a program by changing it, they then need information that tells them how to improve it.
4. A program results in an action in a social environment; it disturbs the system.[5] Studies of this action can yield information about the social problem it addresses, its delivery system, or its own characteristics.
5. Once programs are launched, numerous factors apparently unrelated to them can influence their outcomes. These factors may open the way to diverse interpretations of outcome data. Program effects are therefore inherently less knowable than the effects of a controlled laboratory experiment, which itself is a kind of program.
6. Federal education programs have a pluralistic and decentralized delivery system. States and localities have the primary responsibility and authority for elementary and secondary education, and often have educational aims that differ from those of the federal government. They also vary among themselves in characteristics that affect program implementation, such as financing and administrative procedures, and school-age populations.
7. Federal education programs vary substantially in what they try to do for whom. Programs may support teacher training, innovation, diffusion of effective educational ideas, or instructional demonstrations. They may promote vocational knowledge, reading, or career education. They aim at counselors, particular classes of teachers, institutions such as libraries, all children, limited-English-speaking children, poor children, migratory children, institutionalized, neglected, or delinquent children, and children in desegregated schools.
8. Since the goals and clientele are often not comparable among federal education programs, trade-offs among them involve criteria of value, not efficiency. Efficiency should dominate the choice among programs only if all of them seek the same goals for the same people. Comparative rules for decisionmaking, such as cost-effectiveness models, are therefore usually not relevant to education programs.
9. Any given federal program represents only a small proportion of a school district's total budget. In fiscal year 1975-76, public elementary and secondary education cost $61.4 billion nationally. Of this amount the federal contribution was only $5.3 billion. Title I, the largest single program, accounted for 34 percent of the

---

[5] As the statistician George Box noted, "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)." (Cited in Gilbert, Light, and Mosteller, 1975)

federal contribution; its funding level was $1.8 billion, or only 3 percent of total national expenditures.

10. Most federal education programs are small. In fiscal year 1977, 48 programs were funded. Four programs (8 percent) were responsible for 70 percent of the budget, and 6 programs for 78 percent. Each of the remaining 42 programs involved from 0.2 percent to 2.9 percent of the federal education budget.

Implications for Evaluating Federal Education Programs

Program evaluations primarily operate as feedback on the status or consequences of a policy. Characteristics of the policy process and programs affect what information is feasible and desirable, and therefore establish some guidance for adequate program evaluations.

Policymakers need a broad range of descriptive information. They need descriptive, or "what is the case," information about the programs they are responsible for. The complex federal system makes it difficult to learn what is going on in an education program. Programs operate at the local level, via decentralized and pluralistic delivery chains, which create a distance between the federal and local levels that can be remedied only with formal information flows such as hearings or descriptive studies.

Federal program responsibilities determine exactly what descriptive information policymakers need: descriptions of program cost, implementation, and outcomes.

Policymakers need interpretive information. As we noted, they usually respond to evidence of poor program performance by changing, not ending, the program. Changing a program requires knowing both its current state and ways to improve it. Policymakers therefore need to know *why* unacceptable outcomes are occurring.

Policymakers need different kinds of information as programs evolve. Since federal education programs pass through many hands before reaching their targets, federal actors have little control over what happens. Clearly, whether and how a program is implemented affects the nature and timing of results. Early in a program, policymakers need to know how it is being carried out, where resources are going, what services are being supported, and who is being served. When program operations stabilize, measuring program outcomes then becomes important.

Policymakers continue to need information about old programs still in existence, because social problems and education programs tend to persist and to evolve.

Policymakers need and can use a multiyear information plan to get feedback on a program. Since programs both persist and evolve, policymakers need regular information on their status. They also use program information for predictable decision points in the policy process. A multiyear plan allows these points to be anticipated and information to be collected according to their schedule.

The information needs of major federal education policymakers should guide what program information is collected. Federal policymakers who can change programs are the major users of program information. Their needs should affect decisions about what program information to collect. In some cases, policymakers in different positions in the policy process need the same information; in others, their different responsibilities produce unique information needs.

Most of the funds for evaluation should be spent on the few programs that absorb the bulk of federal education funds. Limited evaluation funds, and limits on the size of the federal evaluation staff, make it impossible to evaluate all programs. The emphasis should be on obtaining information about the six to eight dominant programs.

Concepts of adequate implementation, the evaluation design, and measures for a program should reflect the fact that programs vary in delivery from site to site. Communities vary in characteristics that affect program implementation. Local school districts also have considerable discretion in program delivery. We can therefore expect the program outcomes to vary from site to site. Evaluation designs and measures should be sensitive to these variations.

DOMINANT EVALUATION STRATEGY

We use the term "evaluation strategy" throughout this and the following subsection, defining it as follows:

> *Definition.* An evaluation strategy is defined as a set of prescriptive statements about: what can be known and what it is important to know *(substance)*; when it should be known *(timing)*; and how it should be learned *(procedure)*.

We argue that the idealized strategy treats an educational program as similar to a treatment in an experiment. Although there are analogies between programs and experimental treatments, we believe that the analogy breaks down fairly quickly and should not be used to evaluate federal education programs.

This section analyzes the parallels between experimental practices and the current strategy for evaluating federal education programs. It also indicates how experimental practices are out of step with characteristics of the policy process and federal programs.

Background

Substantial federal involvement in schooling dates from the 1965 Elementary and Secondary Education Act (ESEA). This act was passed in a decade marked by federal attempts to rationalize budget allocations with cost-effectiveness models. The language of early legislative mandates for educational evaluation reflects these models. For example, Title I of ESEA requires states to adopt:

> . . . *effective* procedures, including provisions for appropriate objective measurements of educational achievement . . . for evaluating at least annually the *effectiveness* of the program in meeting the special educational needs of educationally deprived children. [Emphasis added][6]

---

[6] U.S. Congress, House Committee on Education and Labor and the Senate Committee on Labor and Public Welfare, "Elementary and Secondary Education Amendments of 1969" (P.L. 91-230), Part D, February 1975, p. 69.

The 1974 Amendment to the General Education Provisions Act had similar language:

> The Secretary of HEW shall transmit to the Committee on Education and Labor of the House of Representatives and the Committee on Labor and Public Welfare of the Senate, an annual report which evaluates the effectiveness of applicable programs in achieving their legislated purposes.[7]

As the 1965 ESEA quotation reveals, Congress also tended to equate program effects with student outcomes for those federal programs targeted on children. Congress may not have meant the equation to be rigorously applied, but the legislative language was specific.

Over time, strategies for evaluating federal education programs emerged. Evaluators have differed in specifying what norms should govern the evaluation process. However, practices and assumptions associated with psychological laboratory experiments have strongly influenced evaluators, not only in their research designs, but also, more profoundly, in how they conceived of the evaluative enterprise. Numerous books and articles espouse or reflect the experimental tradition.[8] The head of OPBE uses experimental language when he talks about evaluating OE programs.[9] Much of the debate about appropriate evaluation strategies revolves around experimental procedures versus their alternatives (e.g., Rossi, 1972; Weiss and Rein, 1972; Edwards and Guttentag, 1975).

It is not surprising that this tradition became the dominant normative influence in education evaluation. It seemed to fit the questions that many Congressmen asked about program effects. Programs could be interpreted as the manipulated variable in laboratory experiments, and controlled experiments contain a powerful logic for establishing causality between intervention and outcomes. At least some members of Congress and some legislative language equated effects with student outcomes. Again, this problem looked like problems encountered in psychology (e.g., learning problems) that were traditionally explored through laboratory experiments.

Although the experimental tradition seemed consistent with the evaluation requirements that Congress had established, we argue that it and the realities of the policy process and federal programs do not match well and that the mismatch matters, *especially for the large federal programs.*

In some cases the mismatch is due to practices specific to experiments in psychology. For example, in psychological experiments, investigators usually expect to observe effects immediately after the intervention is introduced and are therefore not predisposed to raise questions about time lags between treatment and effects. In certain biological experiments, by contrast, effects are not expected and therefore not assessed for several days or weeks. In measuring student effects, evaluators trained in psychology are predisposed to treat students at different geographic sites as members of the same classroom, or as comparable subjects in a laboratory learning experiment. Agriculturalists are more alert to how intersite variations affect "subject" response as well as the treatment itself.

---

[7] Comptroller General of the United States, *Annual Evaluation Report of Programs Administered by the U.S. Office of Education, Fiscal Year 1976*, General Accounting Office, Washington, D.C., September 8, 1977, p. 1.

[8] Some obvious titles include *Evaluation and Experiment*, Bennett and Lumsdaine, 1975; *Social Experimentation*, Riecken and Boruch, 1975; and *Quasi-Experimental Approaches: Testing Theory and Evaluating Policy*, Caporaso and Roos, 1973.

[9] For example, in a September 1976 interview between a MITRE Corporation representative and Dr. John Evans, Assistant Commissioner for the main evaluation unit in OE, the Office of Planning, Budget, and Evaluation, Evans notes that "... our evaluation model is the classic model of experimental design. If we could, we would have every evaluation carried out in a manner by which we randomly assign projects or individuals to a control or a treatment population and then carry out a longitudinal evaluation of the two groups. That's our basic ideal." (*Proceedings of a Symposium on the Use of Evaluation by Federal Agencies*, Vol. I, App. II, 1976, p. 18.)

In other cases, practices associated with the experimental tradition, regardless of discipline, produce the mismatch. Experiments rely on controlled conditions, whether in physics, biology, or psychology. Evaluators certainly do not treat a federal program as a controlled laboratory experiment. The evaluation literature contains extensive discussion about the quasi-experimental nature of the evaluation problem. However, evaluators tend to treat evaluation problems as deviations from the controlled laboratory case and "patch" on that presumption.

Our quarrel is not with the rigor and system that the experimental tradition has introduced into education evaluation. The tradition has disciplined standards of evidence and inference. Although we believe that these standards have had *substantively* perverse effects on evaluations of the large education programs, the tradition's standards *per se* and discipline *per se* have represented positive influences.

We also suggest that the experimental tradition yields an appropriate strategy for evaluating some programs—for example, smaller project-grant programs in which individual projects include well-defined treatments and objectives. In these cases, policymakers' information needs are closer to those naturally addressed by the experimental tradition. The projects are also more apt to approximate the conditions under which the tradition efficiently eliminates major competing explanations of outcome (e.g., Gilbert, Light, and Mosteller, 1975; Boruch and Wartman, 1979).

However, we think that: (1) the experimental tradition does not adequately handle the evaluation problems posed by any of the large OE programs; and (2) it therefore cannot give rise to an appropriate *general strategy* for assessing federal education programs. For the large OE programs, we have to move to a more eclectic approach that combines elements of several knowledge-producing traditions, such as case studies, and the economist's strategy of looking for a match between theoretical expectations and data and case studies, as well as the experimental tradition.

Although evaluation practice by no means reflects only the experimental influence, the experimental tradition has been the most pervasive influence and warrants the following critique.

The Experimental Tradition and Program Realities

We argue that the experimental tradition is an inadequate and, in some cases, the wrong basis for determining the content, timing, and method of evaluation of federal education programs, for interpreting data, and for defining the evaluative task. We discuss this critique in five parts:

1. Questions asked and not asked;
2. Inferences made from effects data;
3. The timing of effects estimates;
4. Measurement assumptions; and
5. Conception of the evaluative task.

Questions Asked and Not Asked. The psychological experimental tradition directs attention to effects, particularly student effects. It does not focus on questions about the needs of individuals and social groups, about the implementation of programs, and about the *feasibility* of obtaining intended effects.

Outcome Questions. Psychological experiments test causality, usually by assessing the effects of a treatment on subjects. This tradition poses two problems for evaluating federal programs. First, it directs attention to outcomes. But the effects of a federal education program, important as they may be, are not the only important measures for federal program decisions. Measures of the delivery process are equally important.

Second, the tradition directs attention to questions about only certain *kinds* of outcomes, usually individual student outcomes. Major federal education programs have many objectives; usually some of them do not involve student effects. In some cases, the program is expected to affect the allocation of educational *inputs* (e.g., funds in the program for School Assistance in Federally Affected Areas; education services in Title I; attention to a class of children in Title I). In other cases, the program is intended to change the *processes* surrounding the delivery of inputs. For example, the Education for All Handicapped Children Act has a due-process objective. A main objective of Title I was to make poor and minority children important clients of school systems that had previously neglected them. This is a process outcome and is independent of poor children's improved achievement, which Title I also targets. An experiment usually treats inputs and processes as independent and mediating variables, respectively, *not* as effects (dependent variables). Thus, the experimental tradition does not automatically lead to an examination of this wider class of outcomes.

Implementation Questions. In the laboratory the initiator of a treatment, the treatment, and recipients of the treatment are in direct relationship. Careful investigators always determine if subjects "received" the treatment in ways intended by the investigator. However, laboratory researchers can quickly proceed to the question of treatment effects, with limited need to worry about whether or not the treatment occurred, what its characteristics were, or who received it.

Federal education programs are delivered through a complex intergovernmental structure. Even in the simplest delivery case (such as project grants for bilingual education), those who pay for the treatment—for example, Congress or OE federal staff—have no direct relationship with the treatment itself or its recipients. Thus, they have no direct knowledge that the treatment is occurring, of what it consists (perhaps whether the content of a Title I course is reading or mathematics), or of who receives it. In such indirect relationships between the policy and the action, the question of implementation—Is anything being delivered, and, if so, what and to whom?—is major and necessarily precedes that of effects.

Need Questions. In the laboratory, subjects are not of interest in themselves. They simply register what happens when some treatment is administered. Thus, the experimental tradition does not lead to questions about the needs of subjects as citizens. However, for federal education programs, the needs of individual citizens or social groups are the basis for legislative action.

Generalizability Questions. In the experimental tradition, generalizability questions take two forms: Do the same results occur each time under the same conditions? Do the results change when initial conditions are systematically varied?

Both of these questions involve questions about the *stability* of results. However, the experimental tradition does not anticipate having to answer these questions, and does not have procedures for answering them, when conditions vary widely and unsystematically.

Some federal programs, such as the USOE Packaging and Dissemination program, involve diffusing practices that are successful at one site to other sites. Doing so requires either that the practice's success be independent of site conditions or that the originating site's conditions occur widely. As Glennan et al. (1978) observe, however, education is peculiarly sensitive to its context. In other words, we cannot expect the success of the practice to be independent of site conditions.

We also cannot expect conditions to be the same across sites. Schools vary in their administration, financing, school-age populations, and community situations. Thus, we can expect the program's characteristics to differ from site to site. Since individual and group responses seem highly sensitive to context (Cronbach, 1974),[10] we can also expect the responses of program recipients to vary as the program varies.

Inferences Made from Effects Data. In a psychology laboratory experiment, the investigator handles the variation among recipients by randomly assigning them to treatment and control conditions. All subjects in each condition receive the same stimulus (either treatment or no treatment) in the same context, and both treatment and control conditions are isolated from other events that might affect response. These three conditions—random assignment, standardized stimulus and context, and isolation from contaminating events—allow the researcher to attribute the results to the treatments.

Federal education programs rarely have comparable treatment and control groups. The programs do not provide a standard treatment in a standard context, and therefore increase the range of client responses. The treatment is frequently confounded with other events that may well affect the outcomes of interest, such as other federal and local programs with similar objectives, or the unintended diffusion of the treatment to the control schools or classrooms.

As a consequence, outcomes of federal education programs and of experiments cannot be interpreted in the same way. In federal programs, outcomes can reflect initial differences between treatment and control groups. Null effects can reflect the large variation in response produced by variations in treatment and context. Effects cannot be clearly attributed to the federal program since they may reflect the influence of other events that take place simultaneously with the program.

Timing of Effects Estimates. In a laboratory experiment, introducing and implementing a treatment are simultaneous. Unless they have theoretical reason to expect effects to lag the treatment, investigators usually measure treatment effects at the end of the treatment. In education, local program implementation trails federal introduction. The usual experimental schedule for measuring effects is not appropriate for federal programs, and provides no criteria for determining a reasonable schedule.

---

[10] Brickell's report (1976) on an Ohio career education program illustrates Cronbach's point.

Measurement Assumptions. The laboratory experiment presumes a common treatment and consequently can use common measures to assess effects. Some federal education programs, such as Title IVC of ESEA, presume and encourage local variation in treatments and thus probably in outcomes. Evaluations of these programs have recognized the lack of basis for common outcome measures.

However, a more subtle problem is posed by programs that presume common outcomes (for example, student achievement in Title I) but still permit local variations in the delivery of services (for example, instruction in reading versus mathematics). When treatment elements differ by site and their effects are assessed by measures applied commonly across sites, the *match* between treatment and outcome measures will tend to vary across sites. In these cases, outcomes may not be picked up by measures applied commonly across sites, even though they meet program intent. The recent NIE study of compensatory education uncovered exactly this situation (National Institute of Education, 1977). The study found that: (1) commonly applied outcome measures significantly mismatched the services actually delivered by the program; and (2) the services delivered usually met federal intent. We do not know how often these mismatches occur, but they imply a measurement problem more complex than that usually assumed for programs expected to produce common outcomes or than exist in experimental situations.

Conception of the Evaluative Task. The experimental tradition is fundamentally a *methodological* or procedural way of knowing outcomes of an intervention. To the extent that this tradition defines the evaluation task, it tends to define it as a procedural, not a theoretical, problem.

In fact, evaluators tend to define their task methodologically. As Cline (1976) notes, evaluation is typically defined as finding out if something happened as the result of Program A. In consequence, evaluators rarely get involved with models of an educational process. They rarely ask why a program works or does not work. They typically assume that defects in methods, rather than ignorance of the program and its content, are the source of uninterpretable evaluation results. In other words, evaluators typically ask what, not why.

This theory-free approach to evaluation means that evaluations are unlikely to meet two information needs of policymakers identified earlier. One of these needs was to know the effects of a program with some reasonable certainty; the second, to have some basis for adjusting the program. We elaborate each of these points.

Inferences about Program Effects

To assess the effects of a program, we must be able to reliably attribute measured results to the program. This is essentially a question of causality. Researchers generally agree to accept causal inferences as valid under either of two circumstances.

- The investigator has followed certain methods that exclude explanations of the results other than the intervention itself; for example, a well-conducted randomized experiment represents such a method. This mechanistic way of determining causality relies more on procedure than on understanding.
- The investigator *understands* the problem, the intervention, and outcomes. In this case, the investigator knows (or believes he or she knows) which factors other than the intervention can affect the outcomes of interest. The investigator can then measure these factors in the actual intervention, attributing outcomes to the intervention either because: (1) there are no perturbing events, or (2) there are such events, and the results can be "corrected" for their effects, either quantitatively or by judgment.

Psychologists traditionally rely on the experimental method to assess causality, although usually for problems that are well defined and already partially understood. Economists tend to rely on the match between the predictions of theoretically justified models and nonexperimental data. As Cain (1975) observes, the negative income tax experiment is an unusual way for economists to estimate effects.

The evaluation literature recognizes that evaluation designs cannot come close to the methodological rigor of laboratory experiments. However, it rarely observes that: (1) laboratory experiments not only have strong methods, but are also usually tied to substantive theories; and (2) substantive knowledge becomes more important when method is weak. To some extent, substantive knowledge can substitute for method. When analysts know more about a program, they can specify which program effects are reasonable to expect. If outcome data fit these predictions, analysts are in a stronger position to attribute these outcomes to the program. To the extent that evaluators define their task as a methodological task rather than as a methodological and theory-construction task, they fail to exploit a fundamental strategy for assessing causality.

Knowing How to Adjust the Program

Evaluation without theory limits the basis not only for assessing causality, but also for deciding how to adjust the program when outcomes are not strongly positive. If effects are negative, or, as usually happens, null or slightly positive,[11] then policymakers need an interpretive framework for deciding how to act on the outcome data. For example, if policymakers do not know what services programs are delivering, they have no basis for choosing between two explanations of null effects. Do null effects result from a discrepancy between the services legislated and those delivered? Or do they result from ineffective services—ineffective at this funding level, or with the current technology, or for these particular clients?

---

[11] Gilbert, Light, and Mosteller (1975) note that social programs rarely yield "slam-bang" effects. In a sample of medical, socio-medical, and social innovations that had been run as randomized field trials, the authors found that 4 percent of the innovations had a definitely harmful effect; 7 percent, a slightly negative effect; 46 percent, a null effect; 21 percent, a slightly positive effect; and 21 percent, a very positive effect. The authors estimated that the sample of studies was biased toward successful innovations; nevertheless, 68 percent had null or very small effects.

Summary

Experimental treatments and large federal education programs differ in purpose and nature and require different assessment strategies. Laboratory experiments seek to confirm expected causal relationships between inputs and outcomes. Conceptually, the tradition therefore tends to equate outcomes and effects. Since education programs can aim at affecting what are independent (input), mediating (process), or dependent (outcome) variables in an experiment, their assessment requires a broader definition of effects than outcomes alone.

Treatments and programs are also delivered in very different systems. As a result, implementation looms as far more pivotal for federal programs than for laboratory experiments. The delivery system for an experimental treatment is direct and centralized; the initiator and deliverer of the treatment are the same. The delivery system for social programs is indirect and decentralized; the initiator and deliverer of programs are not the same. By virtue of direct contact with the treatment, an experimenter has some intuitive knowledge about treatment implementation; evaluators of federal programs do not. An experimenter can standardize the treatment; the decentralization of a program means that the program delivered varies from site to site. Since the program delivery system is unique across sites, the experimental practice of using common measures across all intervention instances does not usually work for education programs.

The experimental system is reproducible and bounded; that is, events that can affect treatment effects cannot "leak" into the system. Method alone represents a reasonable basis for assessing causality. The program system is loosely bounded: Events affecting program effects can leak into the program delivery system. The program itself can leak out to "contaminate" the programs being used as controls for estimating program effects. Therefore, method is not a reliable way to establish causality for programs, and interpretive frameworks become more necessary for program evaluation than for laboratory experiments.

A preceding section of this essay described what selected characteristics of the policy process and of education programs imply for evaluation. These implications do not match those of the experimental tradition for the large federal programs. The experimental model restricts the range of information collected; being essentially methodological, it does not encourage the development of interpretive frameworks; it does not lead to the idea of sequenced information; and it is antithetical to measures that are consistent with intersite variation. The strategy we propose argues for a substantively expanded range of studies, a sequencing of functionally different studies according to the program's natural life cycle, and, especially for large education programs, a multiyear plan of limited objective studies that are negotiated with major federal users of evaluations, including the Congress.

AN ALTERNATIVE STRATEGY FOR EVALUATING EDUCATION PROGRAMS

This section recommends changes in the dominant evaluation strategy according to changes in *substance, timing,* and *procedure.* It prefaces these recommendations with descriptions of two evaluation programs that substantially influenced our thinking.

Sources of Recommendations

The analysis in the previous sections drew from a variety of sources, but this section is inspired primarily by the National Institute of Education (NIE) Compensatory Education study and the Bureau for the Education of the Handicapped (BEH) plan for evaluating the Education for All Handicapped Children Act (EHA).

NIE Compensatory Education Study.[12] Designating itself as the client, the Congress commissioned the NIE Compensatory Education study in anticipation of the ESEA Title I reauthorization hearings. The study had several features that particularly influenced our thinking: (1) the participation of policymakers in the selection of questions to be investigated; (2) the concept of a large set of individual studies, each with limited objectives; (3) a selection of studies that could yield descriptive information much broader than student outcomes; (4) an explicit strategy for ensuring that study results would be available in time for the Title I reauthorization hearings; and (5) a format for reporting study results that allowed synthesis of results from many studies.

Congress directed the NIE staff to submit a study plan before proceeding further. In developing the plan, the NIE staff consulted with Title I program staff, advisory committees, school administrators, and researchers. Despite this wide consultation, the study director saw himself as serving a select clientele, specifically, the majority and minority staffs of the House and Senate Education Committees. Thus, the primary planning negotiations were between the NIE and Congressional staffs in order to determine what information both the majority and minority staffs could agree would be important in the prospective debate.

On the basis of these negotiations, the NIE staff commissioned 35 separate research projects. The large number of projects resulted from an attempt to limit the objectives of any one, even though this strategy multiplied NIE's contractual problems. The staff expected two benefits from the limited objectives of each project. First, each project was simpler to carry out and therefore more likely to be completed on schedule. Second, if one project failed, it did little harm to the entire study.

The studies commissioned by NIE reflected Congressional needs for a wide range of descriptive information. The studies included such topics as: (1) How did different elements of the funding formula affect the payout of federal funds among districts? (2) How were noninstructional services selected and delivered in Title I schools? (3) What differences were there in how quickly Title I children learned in individualized versus group instructional programs?

The NIE staff treated the Congressional deadlines for receiving the study as fixed, and ensured that results would be available in time for the reauthorization hearings. For example, the NIE staff recognized that producing the final report consumed a substantial part of a contractor's time; consequently, they frequently told contractors what data tables they needed by which date, without requiring written text. The NIE staff assembled the data in their own reports to Congress.

---

[12] Chapter 4 below describes this study in greater detail.

The NIE staff did not report the results of each project individually. The original 35 projects ended up as six synthesized reports to the Congress, establishing one basis for the Title I debate.

BEH Evaluation Plan. The BEH evaluation unit has developed a plan for assessing the Education for All Handicapped Children Act of 1976 (EHA). In some respects the BEH plan is more relevant than the NIE study for evaluating federal education programs. The NIE staff was specially assembled, it had an unusual direct relationship with the Congress, and the study was conducted a decade after Title I began. The BEH evaluation unit operates within the normal political constraints of any HEW unit. The EHA program is new, and an evaluation plan can therefore be designed to assess the program as it is carried out.

Two features of the BEH plan are the same as those of the NIE study: (1) negotiation of the plan with policymakers and interest groups; and (2) the broad range of substantive questions to be investigated. Two others are unique: (3) the concept of a multiyear plan; and (4) the concept of sequencing studies according to a program's natural life cycle.

The BEH staff negotiated both internally (with representatives from the Office of the Deputy Commissioner, the Division of Assistance to States, and the Division of Innovation and Development), and externally (with the relevant Congressional staff; the major organized constituency for handicapped children, the Council for Exceptional Children; the Committee on Evaluation and Information Systems of the Chief State School Officers organization; and the National Association of State Directors and Special Education, an association of those state officials responsible for implementing the law). After each external review, the BEH evaluation staff prepared a written summary of the comments received and sent it back to the participants to make sure that their intent had been understood.

The questions that emerged in the negotiation process show the same broad range as the NIE questions. The final plan consists of six major questions (each further analyzed into subquestions). Who are the intended beneficiaries? Where are the beneficiaries being served? What services are being provided to children? What administrative mechanisms are in place? To what extent is the intent of the law being met? What are the consequences of implementation? They range from questions about the target group, to implementation, to intended effects of the law, and to unintended consequences of the law.

In devising the EHA plan, the BEH staff worked from certain explicit assumptions. One was that the plan should consist of a number of "interlocking" studies conducted over time. The plan in fact consists of a *system* of studies. The other was that these studies should be sequenced according to the natural development of the program. For example, implementation problems are more apt to occur early in a program's life; the BEH staff scheduled studies of EHA's implementation accordingly. This timing gives federal program managers the chance to respond to implementation difficulties as they arise. It also lets the evaluation staff monitor program operations to determine when they can reasonably assess program consequences.

Recommended Changes in the Evaluation Strategy for Education Programs

Based on the foregoing discussion, we propose changes in the strategy used to evaluate the major federal education programs. This section focuses on three subjects: the substance of the studies to be conducted, their timing, and the procedures by which they are defined.

Substance. Our analysis suggests four criteria for selecting studies.

1. *The information that policymakers need should govern the studies selected.* Social program evaluation started when cost-effectiveness analysis was popular among policymakers. The function of program evaluation, not surprisingly, therefore came to be defined as "placing value on" and led to an emphasis on program effects. The NIE and BEH experiences suggest that some policymakers want evaluation to go beyond the measurement of effects.

The traditional functional concept of evaluation seems inappropriate for several reasons. First, the range of studies that it implies is much narrower than what some policymakers need, as the NIE and BEH negotiations suggest.

Second, the concept of "placing value on" presumes that evaluation results can play a role for which they are not well suited. Any single federal education program has many noncomparable objectives that enter into policy decisions about its worth. Education programs also differ too much among themselves in intent and target to allow comparisons along the same dimensions. Disparity of objectives means that no one analytic framework can yield a final judgment about the worth of a program, either singly or in comparison with other programs. Under these conditions, decisions about worth necessarily become political. Evaluation studies can inform political judgments, but cannot prescribe them.

Third, program effects alone do not provide enough basis for interpretation of effects data or for the several policy decisions that can be made about a particular program.

2. *Studies should assess the reasonableness of major assumptions underlying the program legislation.* Such legislation assumes that certain needs exist and that certain objectives are attainable, but is usually based on inadequate knowledge. The enactment of legislation is ordinarily the beginning, not the culmination, of public learning about problems and solutions. At the start, Congress may not accurately estimate the nature or level of need. More important, the required "technology" may not exist, or may be too weak or its effects too sensitive to variations in local conditions to realize particular intents of legislation.[13]

The executive branch is expected to implement Congressional intent. Agency personnel and evaluators therefore usually "accept" Congressional assumptions and "keep trying," an orientation that discourages uneasy questions about the feasibility of the legislation. The scanty knowledge available at the start of most programs argues that some studies should explicitly treat program assumptions as hypotheses to be tested, not as facts. For example, before investigating to see whether a program is achieving Congressional intent, it is reasonable to ask whether the intent is feasible. Evaluating program outcomes under optimal conditions, such as controlled laboratory or exemplary field conditions, shows what the available technology can accomplish at best. The results of these studies indicate whether there is any reason to expect a particular effect under average field conditions.

---

[13] For example, the Packaging and Dissemination program presumes that practices successful in some school districts can be packaged, disseminated to, and successfully incorporated by other districts. However, accumulating evidence on how schools adopt new practices suggests that the program is based on a faulty assumption.

3. *The studies should produce an increasing amount of interpretive information about how parts of the program* (funding and service delivery systems, intended beneficiaries) *interact to produce observed outcomes.*

An interpretive framework[14] emerges if the program's evaluation unit focuses on why, as well as what. An interpretive focus should affect what studies are selected and how any given study is conducted. For example, a study of how program clients use program services can explain program outcomes by revealing the incentive structure that the program offers to clients.

4. *Studies should establish how diverse the program is.* The uniformity or diversity of a program is an empirical question. Some authorizing statutes presume that the program will vary from site to site and thus implicitly anticipate that program outcomes will also vary. Even when the statutes appear to set uniform goals (such as improving the educational performance of poor children), Congress frequently states its intent so generally that it allows the services delivered to vary widely (such as instruction in reading versus mathematics). Since program legislation and regulations permit variation in services, the nature of the American public education system will almost certainly result in program variation from site to site.

Evaluators should document the nature and extent of service diversity for at least two reasons. First, such studies inform policymakers about the reality of the program. Second, evaluators cannot plan intelligent studies of effects without knowing about the variation in services delivered. If a program is supposed to deliver uniform services, using the same measures of outcomes across sites is logically valid and defensible. However, if outcome measures presume uniform services and those delivered are diverse *and* responsive to the intent of the program legislation, program effects will be underestimated.

When the nature of services delivered is unknown, uniform outcome measures are legitimate under only a few conditions. One such condition is when: (1) the legislation seeks effects for a single domain, such as reading, and (2) achievement in the domain consists of a generalized competency that clients can display without having previously encountered specific test items, such as specific reading passages. In this case, site variation in specific lessons is not expected to affect the acquisition of the competency. A second condition is when the legislation offers specific criteria for measuring the program's success. For example, a program could aim at increasing the number of children capable of attaining a minimum score on a given achievement test.

---

[14] The terms "interpretive information" or "interpretive framework" are deliberately used instead of such terms as "theory" or "model." Theories and interpretive frameworks perform some of the same functions: organizing explanatory information, limiting the number of plausible interpretations, and establishing priorities. However, as we use the terms, they differ in important ways. Relative to a theory, an interpretive framework will be less deterministic and less internally consistent. The terms "theory" and "model" imply a delimited domain. An interpretive framework brings together information on diverse phenomena. For example, a framework for interpreting a Title I outcome—children's reading scores—might bring together information on such different phenomena as delivery systems, classroom processes, and children's homes. A good theory will contain all of the statements required to explain phenomena that fall within its scope. Even a good interpretive framework will have large holes in it. A good theory will not yield alternative interpretations of a phenomenon. A framework will yield alternative interpretations, although it will limit the number that seem plausible.

Recommended Range of Studies. These four criteria imply an evaluation strategy with the following kinds of studies:

- Studies of the needs of target populations (e.g., students, teachers) or institutions (e.g., school districts).
- Analyses of the process of implementing a new program or changing an old one.
- Descriptions of the distributions of resources (e.g., funds, services) among school districts or target populations. When the program aims at a certain distribution of resources, these studies are studies of effects. When the desired distribution is a means to an end, the studies are of implementation.
- Feasibility studies to establish the effects that can be expected from a well-implemented program. These studies determine whether the program can meet its objectives, at least under optimal conditions.
- Analyses of the *expected* consequences of initiating, changing, or terminating a program (e.g., likely effects of changes in program regulations).
- Studies of program effects on target institutions or populations (e.g., student achievement).
- Syntheses of knowledge about a particular program for policymakers who are reviewing the program.

Three things should be noted about this array. First, measuring effects remains a legitimate evaluation activity.

Second, we recognize that the studies we recommend include ones not normally considered evaluations. Narrowly construed, "evaluation" refers to effects-studies, although such concepts as "formative evaluation" have extended the concept of evaluation to include implementation studies. Rarely, however, does evaluation refer to needs and feasibility studies. Our intent is not so much to change the definition of "evaluation" as to extend the range of studies that sponsors of evaluations undertake. We want these units to sponsor any studies needed for programmatic decisions, whether or not the studies would normally be called evaluations. From this perspective, perhaps we should use a phrase like "program-related studies" instead of "evaluations." However, since we began this paper as an examination of evaluation policies, we continue to use the term "evaluation studies."

Finally, evaluating a program does not mean that all of these kinds of studies have to be conducted. The information needs of certain decisionmakers in the Congress and the executive branch should determine the studies actually done for a particular program.

Supplementary Program Data. In addition to studies of individual programs, three other kinds of systematic knowledge are relevant to educational programs: (1) synthesis of knowledge from several programs; (2) data from the National Center for Educational Statistics (NCES); and (3) study results from the National Institute of Education (NIE).

The kinds of studies listed earlier refer to individual programs. When studies of several programs are put together, they form a potential source of information about questions not necessarily explicitly addressed by any one study or elucidated by any one program. Federal education programs cover a variety of experiences with the educational delivery system, different governance mechanisms, and different program objectives. Studies of them can yield information on questions such as:

- The limits of federal power to change education at local levels.
- The consequences of different infrastructures for delivering federal education programs.
- The implementation consequences of different types of federal regulations.
- The nature of innovation and diffusion processes in the educational system.
- The limits of education in producing change.

The NCES and NIE are not primarily organized to collect data about education programs. However, the NCES publishes time series data about inputs to the educational delivery system, the delivery system itself, its outputs, and its recipient populations. These data describe the context of federal education programs. NIE provides research on basic learning processes. It also does research and development on: (1) the education system's organization and management; (2) relationships between educational institutions and other institutions; and (3) dissemination networks.

These kinds of knowledge contribute to education program decisions in two ways. First, they provide a basis for interpreting evaluation data on individual programs. For example, the NIE program on Local Problem-Solving provides information on innovation processes in local schools. Policymakers can use this information to decide what they can reasonably expect from federal programs that rely on local innovation for their success. Second, these sources can contribute to initiating or changing program legislation—for example, by identifying changes in the number of children in certain categories. These classes of knowledge could be more fruitfully exploited for program decisions, although we do not address the issue here.

Timing. Programs have a *life cycle* that moves from legislation to implementation to routine operation. Even after a program is well established, it changes over time, although much more slowly than during its implementation. *We recommend that studies of a program be timed to accord with the program's place in its life cycle.*

For example, evaluators should assess program implementation in the program's early years, before they measure program effects. They should estimate program effects only when data indicate that local program operations are reasonably stable. They should estimate effects under *average* field conditions only after assessing them under optimal conditions.

Programs will vary in their life cycles. All other things being equal, conflict between federal and local priorities should extend the time—perhaps indefinitely—required to implement a program. A long administrative chain should require longer implementation times than a short one. For example, a formula grant program administered through State Education Agencies (SEAs) should take longer to implement than a project grants program administered directly through school districts. A program with complex, inconsistent, or vague regulations should take longer to implement than one with simple, consistent, and specific regulations.

Finally, a program that requires more organizational changes at federal, state, or local levels should take longer to implement.

Procedure. Thus far we have argued for a wide range of descriptive and interpretive studies, sequenced to accord with the natural life cycle of the program. This section is about procedures—for example, how to determine policymakers' information needs. The basic concept in this section is that of a *multiyear plan of evaluation studies* negotiated among major policymakers.

1. *The OE pattern of reserving evaluation resources primarily for larger programs should continue.* Although OE had jurisdiction over 48 preschool, elementary, and secondary education programs as of the fiscal year 1977 *Annual Evaluation Report,* four of these programs absorbed 70 percent of the budget. None of the remaining programs accounted for more than 3-1/2 percent. Obviously, any evaluation unit has limited funds and staff. The present policy therefore makes sense.

Smaller programs should be evaluated under particular circumstances. If a constituency for a small program is pressuring the Congress to expand the program, information on the program's performance helps the Congress to assess its worth. In other cases, evaluating a small program can illuminate a larger question; for example, it might reveal fundamental characteristics of the delivery system. In most cases, however, information about the smaller programs will have to come from less formal sources, such as the testimony of interested parties.

2. *Large education programs should have a multiyear plan of evaluation studies.* Large programs such as Title I, Vocational Education, and the Education for All Handicapped Children programs have multiple objectives, complex delivery chains, and many sites. They consequently have complex information needs, and call for a *plan* of studies to be undertaken over the next several years. The governance and intent of the program should determine the questions that these studies should address. Their schedule will depend on the expected life cycle of the program.

3. *The plan of studies should consist of studies with limited objectives, instead of ones with multiple purposes.* We recommend against multipurpose or omnibus studies for three reasons. First, because they are complex they take a long time to plan and to complete. Any evaluation study, limited or not, incorporates a vision of the policy problem. The longer the study takes, the more likely it is that the policy community will reconceive the problem during the study. Omnibus studies thus often become outdated, and their results irrelevant.

Second, if a multipurpose study fails for any reason, the consequent information gaps are large and consequently less remediable.

Finally, a multipurpose study often has conflicting design and management requirements. Resolving the conflicts tends to compromise the quality of the study.

4. *The evaluation plan for an education program should be formally negotiated; the Congress should be a party to the negotiations; and this process should be regarded as an important policy activity.* We have recommended developing a multiyear evaluation plan for each major education program. Here we recommend that the evaluation staff and major federal users of the evaluation results *negotiate* the plan and that they revise it annually by the same negotiation process. The

Congress should be one of the parties to this negotiation.[15] Only Congress can legitimately make final decisions concerning the allocation of resources among competing political interests. Therefore, only the Congress can make many of the major decisions about education programs. Because it also has the major responsibility for establishing the programs' basic goals, Congress's information needs should be carefully considered in preparing the plan of studies.

The idea of a formally negotiated, multiyear evaluation plan is novel. Federal evaluation units arose in connection with attempts to install program budgeting systems in the 1960s. Their organizational independence was protected, because they were expected to provide independent information about program performance to the budget decision process. They were usually placed outside of individual programs, frequently in planning or budgeting operations. They were assured an independent budget either through direct appropriations or by receiving a fixed proportion of the program funds. The major evaluation unit for OE programs, OPBE, was organized this way.

OPBE produces an annual evaluation plan in response to the yearly request of the Assistant Secretary for Planning and Evaluation (ASPE) for evaluation plans for major HEW units. The process begins in the spring and terminates in the winter with ASPE's approval of the plan. OPBE tends to make the major decisions about what is to be evaluated and how the evaluation is to take place. Staffs of some OE programs, usually the larger ones, can influence the annual plan by forwarding their evaluation priorities to OPBE. However, interviewed program staff members report little negotiation of conflicts among the priorities of different programs, at least at the program level. Thus, a program manager who disagrees with the intent of an evaluation may not be able to influence the plan, partly because the feedback process is not fully developed.

ASPE and other elements of the Secretary's office enter the process late—at the stage of a completed draft, and more in the role of reviewers than participants in deciding priorities. The timing of ASPE's entry into the planning process has limited its effects on the basic shape of the annual OPBE plan. Historically, ASPE's use of its input opportunities has varied, depending on the interests of the senior staff.

There has been no clear or regular process for eliciting the concerns of the Congress. Executive departments are reluctant to allow direct, intensive relationships between departmental units and the Congress, and higher-level offices in OE and in the Office of the Secretary tend to mediate discussions between OPBE and Congress. Congress will not be included in decisions about evaluation priorities unless the HEW Secretary's office explicitly supports a direct relationship.

Federal policymakers therefore have had little influence on the annual evaluation priorities because there is no formal tradition of negotiating them; OPBE controls the evaluation budget; and OPBE operates within a tradition of independence.

---

[15] Three experienced members of the executive branch who read our original paper questioned the political feasibility of negotiating such a plan, especially with Congress as a party. We share some of their skepticism, but remain convinced that evaluations will be useful to the policy process only if they address the information needs of the major policy actors. To that end, the planning process should include both executive and Congressional actors. The final section of this essay recommends organizational changes that may be necessary for such a process to work.

Earlier, we listed several complaints about evaluations of federal education programs. Some of these, especially the GAO complaints about the timing, relevance, and communication of results, seem to stem from OPBE's independence from the users of their output. At the same time, the fears that initially led to an independent OPBE cannot be disregarded. Choice of data and methods must be free of major bias if the study is to be trusted and used.

Under some circumstances a *negotiated* plan of studies can resolve the complaints without jeopardizing the integrity of the study. We know from research on the use of policy studies that participation in research planning increases its relevance to the participants. Shifting from one or two large studies to a program of studies, most of which are smaller than current practice countenances, should facilitate this participation. Studies can be individually tailored to deal with narrower and more specific questions. If one party to the negotiation has an interest in a problem that is unimportant to the other parties, its interests can be accommodated without modifying or complicating other studies.

Negotiation may also improve the timing of studies. A program connects with several cycles: the executive and Congressional budget cycles, Congressional reauthorization cycles, and its own life cycle. A negotiated plan allows the parties to establish common expectations about the nature of the program's life cycle. It also enables them to deal with conflicts between the program's "schedule" and their needs for information for budget and reauthorization decisions. Since negotiations on a program plan would presumably occur each year, they would offer a formal opportunity to modify life cycle expectations.

Negotiation will be more effective if the parties have different interests and sources of power, thus creating a healthy adversary situation that protects the integrity of the process. For example, representatives of the Republican and Democratic parties negotiated the NIE Compensatory Education Study plan, and thereby, according to the study director, reduced the likelihood of bias.

While the BEH and NIE studies that we have described involved negotiations concerning the study content, we know of no sustained effort to create and modify study agendas in this fashion. Moreover, it is an open question whether Congressional staff will be willing to take the process seriously; we therefore do not know how these negotiations might proceed and be terminated. It is clear, however, that the head of the evaluation office must be empowered to conclude the negotiations and proceed even if total agreement has not been reached; otherwise, those who wish to avoid evaluation could do so by simply failing to agree. As a practical reality, the outcome of the process will surely depend upon the skill and intent of the head of the evaluation office.

5. *The evaluation unit should synthesize the results of studies in terms of specific policy issues.* The NIE Compensatory Education reports chose not to present the results of individual studies, but instead synthesized a number of studies that related to a topic of interest to the client, Congress. The perspective of the reports was more that of the client than of the evaluator, and their formats were more like that of a good staff report than of a conventional research paper.

One interpretation of complaints about past education evaluations is that the Education Division too seldom presents results in this form. OPBE has occasionally prepared "policy implications memoranda"[16] that draw out the implications of an evaluation study. They seem to be well written, policy-oriented documents. Unfortunately, they have not appeared frequently.[17]

6. *Although syntheses of study results should state the implications of the data for policy alternatives, they should not contain policy recommendations or proposals.* Such recommendations normally involve value judgments and properly fall in the province of policymaking units, not knowledge-producing units.

7. *The plan of studies should include both the individual studies and the major syntheses to be prepared for policymakers.* It should identify: (1) the major issues that policymakers want treated, and (2) the topics and timing for the policy-related syntheses that the evaluation program will produce from the research studies. The knowledge required for the syntheses will determine the individual research projects.

8. *The evaluation staff should brief parties to the negotiation process on the content of the syntheses.* Doing so disseminates the evaluation results by establishing a time for users to focus on study findings. It lets users question the evaluation staff on points of particular interest and also serves to close the evaluation process that the participants initiated.

IMPLEMENTING THE EVALUATION STRATEGY

Our recommended evaluation strategy deviates in several respects from current practice. We tried to design it to yield politically disinterested evaluations of technically high quality and relevance to policymakers' needs. The organizational changes that we recommend are intended to promote that objective. "Technically high quality" implies appropriate substance and method, as defined by the evaluation question and frontiers of the art. "Disinterested" implies an evaluation that is not biased to protect the political interests of the organization that controls the evaluation unit. (No study or evaluation, of course, can escape being influenced by the investigator's personal experience and the historical context in which it is produced.) "Relevance" implies evaluations that address the policy issues of concern to the major users.

Implementing the strategy requires a redistribution of educational evaluation resources and may require additional resources. Two of our recommendations will entail more staff time: the negotiation-synthesis process, and limited-purpose studies. A systematic negotiation-synthesis cycle will take more time than OPBE now spends on negotiations with users and on Policy Implications Memoranda. Several limited-purpose studies will also consume more contracting and monitoring time than will one multipurpose study, even if they address no additional questions. The negotiation-synthesis cycle will also require more *user* time and attention from the users of evaluations.

---

[16] U.S. Department of Health, Education, and Welfare Memo to the Secretary from the Commissioner of Education (Policy Implications Memorandum concerning the OE/OPBE Study entitled, *An Analysis of the Relationship Between Reading and Mathematics Achievement Gains and Per-Pupil Expenditures in Title I Projects, Fiscal Year 1972*), October 23, 1973; and U.S. Department of Health, Education, and Welfare Memo to the Acting Commissioner of Education from the Assistant Commissioner for Planning, Budgeting and Evaluation (Policy Implications Memorandum—based on the Study by the American Institutes for Research entitled, *The Identification and Description of Exemplary Bilingual Education Programs*), September 14, 1976.

[17] *The Annual Evaluation Report on Programs Administered by the U.S. Office of Education* combines a history of each OE program with a brief synthesis of research on the program's effectiveness. The annual report is too condensed to serve as a data source for detailed research, but it is an admirable reference document.

We organize our implementation ideas according to the criteria of quality, disinterest, and relevance.

Technical Quality

The technical quality of evaluations normally depends on the competence of staff members and the skill mix available within the evaluation unit.

We make two recommendations on this subject. First, *we recommend a centralized unit for evaluating federal education programs.* Currently, the evaluation function is largely centralized in OPBE, except for the work of the Division of Innovation and Development of the Bureau of the Educationally Handicapped. (Special-purpose and short-term staffs have conducted the Congressionally mandated NIE evaluations.) Centralization contrasts with the practices of other HEW units. In the past, for example, health and human resource programs have conducted their own evaluations.

Centralization will allow a larger staff, which will enhance technical quality in three ways: (1) Being more visible, it can attract better professionals; (2) it has more positions for a variety of specialized skills; and (3) the larger variety and number of jobs facilitates staff development by enabling its members to widen their versatility in related tasks, and to move into higher positions as their experience and skills increase.

We also recommend that *the unit director be a presidential appointee, subject to Senate confirmation,* thereby investing the unit with the higher status that attracts higher-quality leaders.

Disinterest

Three main factors affect the disinterest of evaluations: (1) the organizational location of the evaluation unit; (2) the unit's method of releasing results to policymakers; (3) and the policymaking community's definition of the unit's mission, as reflected in the qualifications expected of a director.

We have recommended that the evaluation function be centralized, not dispersed to the programs. Associating evaluation with individual program units can easily lead to bias in favor of the program because of the natural tendency to protect and advance the interests of the program. Centralizing evaluations in an independent unit protects evaluation from this potential source of bias.

We recommend that the evaluation unit be legislatively established as a unit that reports directly to the Secretary of the Department of Education. It would then be an independent unit within the Department. For the same reasons that we recommend against decentralizing evaluation, we also recommend against placing the evaluation function in agency-wide policy-planning units. Heads of such groups (e.g., ASPE) serve as policy advocates for the Secretary and the administration in power, thereby tending to favor findings that support administration positions. The political position of these units is not compatible with the impartiality that evaluation requires. The legislative mandate for the evaluation unit would heighten its sense of political independence.

We recommend that the head of the evaluation unit report evaluation results simultaneously to executive and Congressional users. This rule has governed the timing of reports by the Bureau of Labor Statistics, such as the monthly release of unemployment statistics. It settles the issue of which user hears first. Information is power, and first hearers can have an advantage over subsequent hearers. Our suggestion does not eliminate the inherent competition among users, but it removes the evaluation unit from this particular power game.

We recommend that the head of the evaluation unit be chosen from applicants of proven professional competence and integrity. Our models for the unit director's "job description" come from the Bureau of Labor Statistics and the National Center for Educational Statistics. Professional norms govern the selection of job candidates for both of those units. These norms help to define, as nonpolitical, the director's position and the unit that he or she directs.

Earlier we recommended that evaluation unit directors be presidential appointees. This makes it easier for presidents to fire incumbents, but the job standards recommended for this and similar positions (e.g., head of NCES) have historically constrained presidents' choices of replacements. They work against the selection of candidates partisan to the President.

Relevance

The relevance of evaluations chiefly depends on two factors: the responsiveness of the evaluation unit to user information needs, and the users' ability to identify their needs and assimilate results.

Organizational arrangements that increase the unit's autonomy work against its responsiveness to major users. Two of our recommendations attempt to redress this imbalance.

The first is our recommendation that the unit director be a presidential appointee, subject to Senate confirmation. (This process includes the Senate hearings and courtesy visits by the candidate to each of the relevant senators prior to the hearings.) That should not only confer high status on the director, but should also increase relevance by stressing the director's joint responsibilities to the Congress and executive branch. During confirmation the Senate committee members can specify what they expect of the evaluation unit.

We also recommend that the *head of the evaluation unit deal directly with relevant Congressional staffs.* Heads of executive departments naturally prefer to limit direct contacts between departmental subunits and the Congress, but the result has been that Congress tends to dismiss evaluation results, often believing that evaluations ignore legislative information needs. The unit will work better if it has free communications with Congress.

Congress is a major audience for program evaluation data, but Congressional staffs believe that they have lacked relevant evaluation data in the past. Consequently, Congress is mandating an increasing number of evaluations to be supervised by Congress. Unless the evaluation unit in the new Department of Education meets Congress' needs better, Congress will probably continue to bypass the Administration by mandating its own evaluations. We may then expect to find more and more resources devoted to competing evaluations. Evaluators' direct access to the Congress does not guarantee more relevant data to that body, but without access the unit cannot give Congress the information it needs.

Finally, *we recommend that executive branch users (such as the new Department's analog to ASPE) establish small analytic staffs to work with the evaluation unit.* These staffs should perform two functions: negotiation and synthesis of policy recommendations. The policy unit staff would identify its information needs and work with the evaluation unit to include them in the annual plan. If policy units treat these annual evaluation negotiations casually, the studies will remain somewhat irrelevant, no matter how the evaluation unit behaves. A policy staff with permanent negotiation responsibilities can bridge the present gap between the evaluation and policy units.

The second function of these analytic staffs would be to synthesize data from the evaluation unit and other research organizations (such as the NCES and NIE) into policy recommendations for the unit. Although the evaluation unit addresses policy *issues*, their syntheses should not include policy recommendations, which properly lie in the province of the policy units. The policy recommendation step increases the chances that relevant information will be used. It connects the evaluating unit's work to actions that the policy unit can take; therefore, the task should be an explicit function of the policy unit's analytic staff, rather than an ad hoc exercise.

We make no similar recommendations for Congressional staffs. In comparison with the executive branch, the legislature depends less on formal organization than on personal contacts for obtaining information. We do suggest that Congressional staffs use their other analytic resources—such as the Congressional Reference Service, the Congressional Budget Office, and the General Accounting Office—to transform research results into policy recommendations.

In sum, we propose a centralized evaluation unit, reporting directly to the Secretary of the Department of Education. The President appoints the director, subject to Senate confirmation. The evaluation unit is nonpolitical, its director being selected on the basis of professional competence and integrity. He or she deals directly with Congressional staffs and reports evaluation results simultaneously to all major executive and Congressional users.

Executive users would have small analytic staffs with two explicit functions: (1) to define information needs and to ensure that the annual negotiated evaluation plan reflects those needs; (2) to assess the policy implications of evaluation syntheses and related information.

We believe that this structure corresponds to the organizational requirements of the evaluation strategy that we recommend. The result should be evaluations that are relevant, of high technical quality, and disinterested.

BIBLIOGRAPHY

Bennett, Carl A., and Arthur A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs*, Academic Press, Inc., New York, 1975.

Berman, Paul, et al., *Federal Programs Supporting Educational Change,* Vols. I-IV, The Rand Corporation, R-1589/1-4, April 1975.

Boruch, Robert F., and Paul M. Wortman, "Implications of Educational Evaluation for Evaluation Policy," in David C. Berliner (ed.), *Review of Research in Education,* Vol. 7, American Research Association, Washington, D.C., 1979, pp. 309-361.

Brickell, Henry M., "Needed: Instruments as Good as Our Eyes," *Journal of Career Education,* Vol. 2, No. 3, Winter 1976, pp. 56-66.

Cain, Glen G., "Regression and Selection Models to Improve Non-Experimental Comparisons," in C. A. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* Academic Press, Inc., New York, 1975.

Campbell, Donald T., *Qualitative Knowing in Action Research,* Occasional Paper of the Stanford Evaluation Consortium, Stanford University, Stanford, California, 1974.

Caporaso, James A., and Leslie L. Roos, Jr. (eds.), *Quasi-Experimental Approaches,* Northwestern University Press, Evanston, Illinois, 1973.

Chelimsky, Eleanor, *An Analysis of the Proceedings of a Symposium on the Use of Evaluation by Federal Agencies, Symposium Report, Vol. II,* The METREK Division, The MITRE Corporation, McLean, Virginia, November 17-18, 1976.

Cline, Marvin G., "The 'What' Without the 'Why' or Evaluation Without Policy Relevance," in C. C. Abt (ed.), *The Evaluation of Social Programs,* Sage Publications, Beverly Hills, California, 1976.

Cohen, David K., and Michael S. Garet, "Reforming Educational Policy with Applied Social Research," *Harvard Educational Review,* Vol. 45, No. 1, February 1975, pp. 17-43.

Coleman, James S., *Policy Research in the Social Sciences,* General Learning Press, Morristown, New Jersey, 1972.

Comptroller General of the United States, *Problems and Needed Improvements in Evaluating Office of Education Programs,* General Accounting Office, Washington, D.C., September 8, 1977.

Cronbach, Lee J., *Beyond the Two Disciplines of Scientific Psychology,* Occasional Paper of the Stanford Evaluation Consortium, Stanford University, Stanford, California, 1974.

Edwards, Ward, and Marcia Guttentag, "Experiments and Evaluations: A Reexamination," in C. A. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* Academic Press, Inc., New York, 1975.

Foskett, William H., "Practical Application of Evaluation in Legislative Processes," *Legislative Oversight and Program Evaluation,* a seminar sponsored by the Congressional Research Service, prepared for the Subcommittee on Oversight Procedures of the Committee on Government Operations of the United States Senate, Washington, D.C., May 1976.

Gilbert, John P., Richard J. Light, and Frederick Mosteller, "Assessing Social Innovations: An Empirical Base for Policy," in C. A. Bennett and A. A. Lumsdaine (eds.), *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs,* Academic Press, Inc., New York, 1975.

Glennan, Thomas K., Jr., et al., *The Role of Demonstrations in Federal R&D Policy*, The Rand Corporation, R-2288-OTA, 1978.

House, Ernest R., et al., "No Simple Answer: Critique of the Follow Through Evaluation," *Harvard Educational Review*, Vol. 48, No. 2, May 1978, pp. 120-160.

Kuhn, Thomas S., "Reflections on My Critics," in I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, London, 1970a.

——, *The Structure of Scientific Revolutions*, The University of Chicago Press, Illinois, 1970b.

Lakatos, Imre, and Alan Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, London, 1970.

Lynn, Laurence E., Jr., "A Federal Evaluation Office?" *Evaluation*, Vol. 1, No. 3, 1973, pp. 56-59.

——, "Policy Relevant Social Research: What Does It Look Like?" in M. Guttentag and S. Saar (eds.), *Evaluation Studies Review Annual*, Vol. 2, 1977.

Menges, Constantine C., *Knowledge and Action: The Use of Social Science Evaluations in Decisions on Equal Educational Opportunity, 1970-72*, paper prepared for the National Institute of Education, Washington, D.C., March 1, 1978.

Morrill, William A., and Walter J. Francis, "Evaluation from the HEW Perspective," unpublished paper presented at the Workshop on Program Management, Federal Executive Institute, Charlottesville, Virginia, May, 1976.

National Institute of Education, *Compensatory Education Services*, U.S. Department of Health, Education, and Welfare, Washington, D.C., July 31, 1977.

Popper, Karl R., *The Poverty of Historicism*, Routledge and Kegan Paul, Limited, London, 1957.

Riecken, Henry W., and Robert F. Boruch, *Social Experimentation*, Academic Press, Inc., New York, 1974.

Rittel, Horst W., and Melvin M. Webber, "Dilemmas in a General Theory of Planning," *Policy Sciences*, Vol. 4, 1973, pp. 155-169.

Rossi, Peter H., "Testing for Success and Failure in Social Action," in Peter H. Rossi and Walter Williams (eds.), *Evaluating Social Programs*, Seminar Press, New York, 1972.

Rossi, Peter H., and Walter Williams (eds.), "Evaluating Social Programs," Seminar Press, New York, 1972.

Timpane, Michael (ed.), *The Federal Interest in Financing Schooling*, Ballinger, Cambridge, Massachusetts, 1978.

U.S. Department of Health, Education, and Welfare, *Annual Evaluation Report of Programs Administered by the U.S. Office of Education, Fiscal Year 1976*, Office of Education, Office of Planning, Budgeting and Evaluation, Washington, D.C., 1976.

——, Division of Innovation and Development, *Evaluation of the Education for All Handicapped Children Act, Public Law 94-142*, Bureau for the Education of the Handicapped, no date.

——, Memo to the Secretary from the Commissioner of Education, OE/OPBE Study entitled, *An Analysis of the Relationship between Reading and Math-*

ematics Achievement Gains and Per-Pupil Expenditures in Title I Projects, Fiscal Year 1972, October 23, 1973.

——, Memo to the Acting Commissioner of Education from the Assistant Commissioner for Planning, Budgeting and Evaluation, Policy Implications Memo—based on the Study by the American Institutes for Research entitled, *The Identification and Description of Exemplary Bilingual Education Programs*, September 14, 1976.

Weiss, Robert S., and Martin Rein, "The Evaluation of Broad-Aim Programs: Difficulties in Exerimental Design and an Alternative," in Carol Weiss (ed.), *Evaluating Action Programs: Readings in Social Action and Education*, Allyn and Bacon Co., Boston, 1972.

Williams, Walter, "The Capacity of Social Science Organizations to Perform Large-Scale Evaluative Research," in Peter H. Rossi and Walter Williams (eds.), *Evaluating Social Programs*, Seminar Press, New York, 1972.

Williams, Walter, and Richard F. Elmore (eds.), *Social Program Implementation*, Academic Press, New York, 1976.

Chapter 3

EVALUATION AND ALCHEMY[1]

by

Milbrey Wallin McLaughlin

Evaluation has become a growth industry, particularly in education. Until the 1960s, "educational evaluation" meant little more than student report cards to most people. Questions of "significant differences" or "cost/benefit" or "treatment effects" were rarely asked. Now, it is safe to say, evaluation is enthroned as the *sine qua non* of scientific management and the cornerstone of accountability. And efforts to assess the effects of federal education policies and special programs carry a hefty price tag. The annual budget of the Office of Planning, Budgeting, and Evaluation within the U.S. Office of Education mushroomed from $1.2 million in 1968 to approximately $21 million in 1977.[2] Congressional mandates to assess local projects funded by federal education programs add millions more to the annual cost of evaluation. But despite the energy and resources devoted to the task, many researchers and practitioners believe these evaluation efforts are largely a waste of time and money.[3]

Rand's study of federally funded "change agent" projects (the so-called Change Agent study) bears directly on this concern.[4] In the course of the study, we looked closely at the local change process as we examined a variety of federal policies, local programs, and school districts in order to understand why some programs are more successful than others. Our research supports the charge that much of the present evaluation is irrelevant and inappropriate—that most evaluations ask the wrong questions and use the wrong measures. Furthermore, the Change Agent study suggests that the problem with contemporary evaluation is deeply rooted—that it stems from a fundamental lack of fit between the evaluation paradigm and the way local projects operate and change takes place. This essay draws on the findings of that study to identify misconceptions in traditional evaluation paradigms and to suggest how we can learn to do better from the experience of special projects.

---

[1] Richard F. Elmore of the University of Washington provided many insights about the problem of evaluation in the course of our discussions on the topic. His comments on an earlier draft were extremely valuable in helping me recast the arguments presented here. I am grateful for his assistance, but absolve him from any responsibility for errors and defects. This paper was originally presented at the August 1978 meeting of the American Political Science Association in New York City.

[2] General Accounting Office, *Problems and Needed Improvements in Evaluating Office of Education Programs*, Report to Congress, Comptroller General of the United States, September 1977.

[3] See, particularly, D. K. Cohen and M. Garet, "Reforming Educational Policy with Applied Social Research," *Harvard Educational Review*, Vol. 45, No. 1, 1975, pp. 17-43; J. David, *Local Uses of Title I Evaluation*, Stanford Research Institute, Menlo Park, California, 1978; J. Murphy, *Doing Qualitative Evaluation*, (forthcoming); C. Weiss, *Evaluation Research*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1972; and idem, *Using Social Research in Public Policy Making*, D. C. Heath, Boston, 1977.

[4] Under the sponsorship of the U.S. Office of Education, Rand conducted a several-year, two-phase study of these projects, which normally offer temporary federal funding to school districts as "seed money." If an innovation is successful, it is assumed that the district will incorporate it and disseminate part or all of the project using other sources of funds. The Rand study analyzed the effects of these federal policies on local change processes.

The first phase of the research (July 1973-April 1975) examined four federal change agent programs (Elementary and Secondary Act Title III, Innovative projects; ESEA Title VII, bilingual projects; Vocational Education Act, 1968 Amendments, Part D, Exemplary Programs; and the Right-to-Read Program). The research addressed issues related to the initiation and implementation of "change agent" projects, looking specifically at what kinds of strategies and conditions promote change and what kinds do not.

The final phase of the research (May 1975-April 1977) examined what happened to local projects in the two largest change agent programs—ESEA Title III and ESEA Title VII—when federal funding stopped. This phase focused on the different forms that local implementation or continuation may take, and analyzed the institutional and project factors that promote or deter the sustaining of Title III and Title VII projects.

The findings of the study are reported in eight volumes under the general title *Federal Programs Supporting Educational Change*, R-1589-HEW, April 1975.

The major findings of the Change Agent study are:

- *Implementation* dominated the outcome of planned change efforts.
- The process of implementation is shaped primarily by *local* factors—rather than by the adoption of a particular technology, the availability of information, funding level, or particular federal program sponsorship.
- Effective implementation is characterized by a process of *mutual adaptation,* in which both the project and participants change over time.

Taken together, these findings mean that the adoption of a "better" practice does not automatically fulfill its promise of "better" outcomes. Initially similar programs installed in different settings undergo unique alteration. What works in one school or school district is not always successful in another one.

We found that new project methods or materials influenced project outcomes *only in interaction* with local institutional and management factors. We found that there was no such thing as a standard or uniform practice; project treatment had only a *second-order* or *indirect* effect on the outcome of locally planned change efforts.

But most project evaluations use an "impact" or "production function" model that correlates initial project inputs, such as level of funding and technology, to project outputs such as student achievement gains. This evaluation approach assumes that project treatment has a primary and direct effect on project outcomes. Local institutional factors and other variables that are important in the implementation process typically are omitted. Evaluators have lightly referred to these local factors as components of the "black box." The contents of the black box, it turns out, matter more to project outcomes than do the other factors that evaluators attempt to calibrate and assess. Insofar as unidentified causal variables have major and direct effects on project outcomes, an evaluation that fails to measure these variables can produce a specious finding of "positive program effects" or of "no significant difference." One result, as program advocates fear, is that estimates of treatment effects will be biased to an unknown degree and that crucial explanatory factors will not be identified or measured. Another result, as many practitioners fear, is that replication of "proven" practices will be urged when those factors that mattered most to the project's success have been neither explored nor explained. Omission of these major "black box" variables precludes reliable or valid assessment of project effects. It also precludes understanding project success or failure in ways that would allow decisionmakers to learn from experience.

The Change Agent study points to a number of factors that are generally ignored in special project evaluations, but that are required for a valid evaluation design. First, we found that institutional support and receptivity to a planned change effort were essential for project success. Administrators' attitudes toward a special project signaled project staff about how seriously they should take project objectives, and thus affected the depth of their commitment. Unless the project seemed to represent a school or district priority, staff usually did not put in the time and effort necessary to make it succeed. Consequently, when administrators showed no support—as was the case when projects were initiated mainly for opportunistic reasons (for example, to reap windfall federal money)—projects were likely to suffer from pro forma implementation or to break down altogether. The result in either case was no "project effect," regardless of project potential or "proven" success elsewhere.

The commitment of project staff is also influenced by the way that project planning, another local variable seldom mentioned, was carried out. We found that successful implementation was almost always preceded by a broad-based planning strategy in which everyone involved was consulted. When teachers who were going to carry out the project also took part in planning, they were more eager to make the project work. Projects initiated at the "grassroots" did not necessarily work out best. We found that it was not the source of an idea that mattered, but *who* was involved in the planning. Staff often resisted projects initiated by the central office as something done "to" them. Conversely, projects developed at the school level, without district administrators' participation, usually floundered because project staff believed that no one cared. They thought it was not worth the effort, in terms of professional self-interest. *Institutional support, receptivity,* and the resulting *staff commitment* are necessary (though not sufficient) conditions for effective implementation. And they are often independent of project promise.

A second local factor affecting implementation is the *capacity* and *expertise* that local staff bring to the project. When the outcomes of special projects are tallied or compared, evaluators seldom ask "relative to what?" Consider, for example, two nominally similar mathematics projects based on Piagetian principles. In one district, project teachers had worked with this method for several years. Project implementation, in this case, involved refining classroom materials. In another district adopting a similar project, teachers had used a different approach to mathematics instruction, which involved staff in fundamental issues of understanding the project philosophy and learning new techniques. At the end of special funding, the skills of the staff in the second project were about equal to those of teachers in the first project as they *began* implementation. What constitutes success in this case and how are project outcomes to be compared? Although teachers in the second project made greater changes, the effect on students was likely to be greater in the first project. Few evaluations consider the important question of baseline capacity either in assessing project impact or in making cross-site comparisons to determine which project "works best."

In addition to local motivation and capacity, the level of project implementation, and thus project outcomes, are influenced by a third factor, *local implementation choices.* Similar projects or "treatments" can be implemented quite differently depending on local decisions about how to put the project into practice. We found that, even where institutional support for the effort was high, these local choices could determine the success or failure of projects, independent of the type of innovation being carried out.

Local choices were particularly important in five areas:

- *Staff Training.* Effective training for project implementation was an iterative process that was informed directly by participant needs and suggestions. Effective training was also concrete, tied to the actual problems confronted by implementors. "One-shot" or rigidly sequenced training, as well as training that turned on abstract discussions of philosophy, did not result in effective change. To this point, effective training typically relied on local resource personnel. Outside consultants, even when they addressed appropriate problems, generally were not there when (usually unpredictable) project information needs arose.
- *Site Selection.* Some districts chose to concentrate a project in a few sites, while others chose many project sites, often in the hope that project effects would be multiplied. Projects that were spread thinly across the district usually failed to get off the ground. A "critical mass" of project participants, providing moral support as well as substantive help, was necessary to support implementation.
- *Local Material Development.* Effective implementation was accompanied by local development of project materials. Sometimes it involved the preparation of project materials from scratch, and sometimes the review and synthesis of commercially available materials. Effective projects did not use "packaged" materials, without at least modifying them. This development activity provided an important learning opportunity for project staff —a chance to think through project concepts in terms of their own classroom, and to "learn by doing."
- *Staff Participation in Project Decisions.* Successfully implemented projects routinely involved staff in making decisions about appropriate modifications to project methods and materials, usually in regular staff meetings. This participation seemed to have two important benefits. Staff commitment was strengthened because regular participation made the project "theirs." At the same time, the suggestions offered by the staff were instrumental in project success. Staff charged with day-to-day responsibility for implementation often are in the best position to identify problems early and to suggest or assess remedies.

Taken together, these local implementation choices made it possible for participants to adapt the project to the reality of the institutional setting, and to modify their practices in response to the project. These strategies supported the process of mutual adaptation that characterized successful projects. Effective implementation strategies provide project staff with necessary and timely feedback, allow project-level corrections, and encourage commitment. For project participants, as well as the broader institutional setting, the process of implementation is heuristic—a process of learning and adjusting, rather than a process of installation.

Finally, project outcomes are often shaped by local events and characteristics unrelated to the project. For one, the outcome of special projects is influenced by such pervasive local characteristics as institutional climate and leadership style. But in addition, the fluid institutional setting constitutes a reality to which planned change efforts must respond and which often deeply affects project implementation. For example, project training may have been deficient, not by design, but because of district cutbacks in resource staff. Or the planning period for an ambitious project may have been compressed because of more urgent issues, such as a teachers' strike or the need to address a new state mandate. Or a critical mass of project staff may have been dispersed by a busing order. In such cases, it is not very useful to know only that training, or planning, or staff concentration was inadequate. Planners and policymakers need to know whether inadequacies resulted from project choices or unrelated factors in the project setting.

These local factors—institutional motivation and capacity and implementation choices, together with normal pressures and institutional characteristics—have strong and direct effects on the outcome of planned change efforts, determining which project techniques are used, where, and how. They also determine whether a promising project strategy "makes a difference," or whether it is implemented at all. Yet these local factors are seldom included in project evaluation models. An analysis that asks how well something is done, before exploring questions of how, or organizational issues of why, is bound to be incomplete and misleading.

If this is the reality of the local change process, what are the implications for project evaluation? At the very least, this view suggests that we need to revise our specification of "treatment." Most evaluations either specify treatments in generic terms—such as the Southwest Laboratory Reading Program or the Bank Street model of early childhood education—or in terms of the original project proposal. Successful projects, however, change significantly over time. Treatment as specified at the outset will probably differ in important respects from treatment in practice as the project nears completion. Failure to track and document these modifications compromises the validity and utility of project evaluations.

This view of the local change process also means that the "project" model adopted by most evaluators is bound to produce misleading results. A special project cannot be validly assessed in isolation from its system context.

A more fundamental implication of this view is that efforts to "fix" existing evaluation paradigms are unlikely to be fruitful. The time-honored experimental model, for one, cannot produce valid findings if the reality described by the Change Agent study is correct. An experimental paradigm assumes control of *all* sources of systematic variance in project effects. These concerns lead evaluators to calibrate carefully such factors as time spent in treatment, teacher experience, socioeconomic background of students, and so on. However, the Change Agent study points to major sources of project variation both within and among adopting systems—such as participant motivation and institutional support—that cannot be controlled by research design. The experimental paradigm, a useful tool in medicine and social psychology, where treatment is discrete and standard, cannot be fixed to fit the reality of planned change in social systems such as public education.

Neither can the input/output or correlational model necessarily be "fixed" simply by introducing quantitative measures of important local or black-box factors. Because there are important qualitative differences in these factors, quantification can mean that essential characteristics are omitted. These omissions have major and direct effects on project implementation and outcome. For example, a mere list of participants in project planning activities does not reveal whether the involvement of important actors has been genuine or only pro forma. Ritualistic participation is no better—and sometimes worse—than no participation at all. Similarly, one cannot simply enter the frequency of project training into an equation. This measure is meaningless unless one knows how useful the training was in the view of the staff. Or, an administrator may express strong commitment to a special project—but more in hopes of mollifying minority parents than in improving classroom services. Such a commitment, however fervent, is not sufficient to motivate project staff.

In short, the Change Agent study suggests that many of the important factors in the local process of change may be inherently unquantifiable and not amenable to control. Our study suggests that in neither the *conduct* of evaluation nor the *utilization* of evaluation results is it possible to structure behavior according to the internal logic of the experimental paradigm. To do so requires the evaluator to become the alchemist, fundamentally transforming elements of the process of change and the business of schooling.

Where, then, does this knowledge of the local setting and the process of implementation lead us? In one sense, it takes us back to square one. What we know about the process of change implies that evaluation models derived from other realities—microeconomics, medicine, and social psychology—simply do not fit the reality of a public social service system, education in particular. The logic of inquiry is wrong. And preoccupation with scientism and with fixing our traditional evaluation paradigms scants what we do know. The general unwillingness of evaluators, funders, project planners, and federal program officials to come to grips with this fundamental incongruence between the set of relationships presumed by our current logic of inquiry and the local reality has led to spending much time and energy in developing new instruments to measure outcome and calibrate inputs. These efforts typically are undertaken at the expense of rethinking the conceptual framework for learning from project experience. Ironically, such tunnel vision can be expected to produce measures that are inaccurate and fundamentally unsound.

One major challenge for evaluators, then, is epistemological: to develop new and valid ways of knowing. There is little precedent in the applied research tradition to aid in the development of an evaluation model that includes organizational and process variables. There is not even consensus about what the important independent variables might be. The Change Agent study and other recent research provide some clues, but we are still some distance from rigorous hypothesis-testing on the process of change.

A second challenge for evaluators and policymakers is to devise different ways to use evaluative information. The findings of the Change Agent study raise serious questions about whether or not evaluation as social accounting is possible at all. If major factors in the local change process are not amenable to control—either in the design of evaluation or the application of evaluation findings—then evaluation cannot serve that purpose. The assumption that there is a one-to-one relationship between what we can find out and what we can affect is not borne out by our research. It is difficult to affect local reality, for the same reasons it is difficult to learn about it. But perhaps it is possible to learn, for example, about the sort of institutional setting that supports effective project implementation. This information could be useful to practitioners and policymakers in deciding which activities to support, where, and with what kind of technical assistance.

A large investment of time and energy in exploratory and largely qualitative research will not appeal to policymakers who have come to take comfort from the "hard science" symbolized by statistical tests. But our evidence suggests that such new approaches are in order if we value learning from experience.

Chapter 4

EVALUATING EDUCATION PROGRAMS FOR FEDERAL POLICYMAKERS: LESSONS FROM THE NIE COMPENSATORY EDUCATION STUDY

by

Paul T. Hill

INTRODUCTION

This essay distills what the author learned about the problems of evaluating federal education programs during three years as Director of the National Institute of Education (NIE) Compensatory Education Study.[1] NIE's was the first of what now appear to be a number of studies mandated by Congress for its own use in reauthorizing federal education programs. This account of the author's experience may therefore be useful to other federal research managers who conduct evaluation studies for Congress; and because virtually all federal program evaluations are at least partly intended for Congressional use, this essay may also help private scholars, research contractors, and federal managers who plan and execute evaluation studies.

The essay is in three parts. The first sketches the background of the NIE study and identifies problems that NIE faced—problems that appear to be common to evaluations of federal education programs. The second part describes how the NIE staff coped with those problems; and because the essay aspires to be more than an interesting collection of "war stories," the concluding section presents the most important lessons to be drawn from the NIE experience.[2]

---

[1] The essay also draws on the shared experience of several NIE researchers. Iris Rotberg and Alison Wolf laid the foundation of the research plan even before the study was mandated by Congress. Iris Rotberg, as deputy director of the study, shared every task, from negotiating with Congress to editing final reports. James Harvey made fundamental contributions to every aspect of the study, and Joy Frechtling, Margot Nyitray, Don Burnes, Ann Milne, and Charles Troob made important contributions to the overall strategy and managed the key projects. They, and many others, are in some sense the coauthors of this paper.

Sue Berryman, Tom Glennan, and John Pincus of Rand, and Jerry Fletcher of the Office of the Assistant Secretary for Education, made valuable comments and suggestions on earlier drafts of this essay.

[2] The following reports are available from the National Institute of Education, 1200 19th Street, N.W., Washington, D.C., 20208:

Evaluating Compensatory Education: An Interim Report on the NIE Compensatory Education Study, December 1976;
Compensatory Education Services, June 1977;
Administration of Compensatory Education, September 1977;
Title I Funds Allocation: The Current Formula, September 1977;
Using Achievement Test Scores to Allocate Title I Funds, September 1977;
Demonstration Studies of Funds Allocation Within School Districts, September 1977;
The Effects of Services on Student Development, September 1977.

THE NIE STUDY AND THE PROBLEMS IT FACED

The NIE Compensatory Education Study was established by Section 821 of PL 93-380, the statute that reauthorized Title I of the Elementary and Secondary Education Act (ESEA) in 1974. That law required NIE to conduct a broad study of ESEA Title I and similar compensatory education programs conducted by several states. The legislative mandate identified social problems for study, required NIE to submit an official research plan for Congressional review, established deadlines for interim and final reports to Congress, and provided a budget of $15 million for the three-year study. Finally, the mandate required NIE to submit its reports directly to Congress without any prior outside review.

By requiring reports in time for the next Title I reauthorization hearings, Congress showed that it wanted to use the results in legislative deliberations. By making Congress the sole judge of NIE's research plan, and the first recipient of its results, Congress stressed that its own interests were to dominate the study.

The mandate also identified some topics for special attention. In particular, it directed NIE to investigate Representative Albert Quie's 1974 proposal to change the statute so as to allocate Title I funds on the basis of counts of low-achieving children rather than low-income ones. Congress defeated that proposal in 1974, but many key members of the House Education Committee felt some sympathy for it. As the ranking member of the House Committee, Mr. Quie was thus able to gain support for measures, such as the NIE study, that would keep his idea alive.

The mandate also asked for "an examination of the fundamental purposes of [compensatory education] programs and the effectiveness of such programs in attaining such purposes," and for an analysis of how well individualized instructional plans worked in the schools.

When the mandate was signed into law in August 1974, it presented a number of unfamiliar problems for NIE. Until then, NIE had never done a study at the behest of Congress, nor had it ever evaluated a major public program. Some researchers on the NIE staff had experience in evaluating compensatory education, but none of them had worked on the noninstructional aspects of a national program.

NIE also had political problems. The Institute was struggling for its life after a series of devastating reverses dealt by the Senate Appropriations Committee. Some NIE supporters in Congress saw the mandated study as NIE's chance to redeem itself; an unspoken implication was that failure of the study might be fatal to the Institute. Other federal agencies had found research on compensatory education to be a minefield. Evaluations of Head Start, Follow Through, and ESEA Title I had produced intense methodological disputes in the research community. These evaluations had generally drawn discouraging conclusions about the effectiveness of compensatory instruction, leading supporters of the program to oppose the whole idea of evaluation. Congressmen interested in education programs did not know what to make of these evaluations, and were disturbed by the resulting controversy. Conducting evaluations had been a major source of political trouble for the U.S. Office of Education. In turning toward NIE for a major evaluation, Congress had imposed some important new political risks on the Institute.

The significance of all this was not lost on the management of NIE. NIE established a special unit to conduct the study, and agreed to give the unit's leaders a free hand in dealing with Congress, interest groups, and researchers. This group, called the NIE study staff, was to plan and manage the research and write the reports to Congress. Thus, for the purpose of the study, the "Institute" was to be the study staff. The rest of NIE might contribute people and ideas at the request of the study director, but they were not involved in routine management.

Once responsibility for the study was clearly delegated, the staff faced five major technical and tactical problems:

1. *How to move from the broad research objectives set by Congress to specific statements of researchable problems.* Only a few of Congress's objectives had implied specific focus and conduct of particular research projects. Most of the objectives required interpretation and refinement in order to be specified as problems suitable for formal research.

2. *How to guarantee that the research made a fair assessment of the strengths and weaknesses of the Title I program.* It was clear that the results of the Title I study could affect how Congress saw the program, and thus its prospects for reauthorization and funding. The only guidance available to NIE at first was the negative example of an evaluation strategy that was widely regarded as incomplete and biased against the program. The outcome measure previously used—student achievement test scores—was under criticism. People believed that it failed to take account of the program's social objectives, and was technically inadequate, even for estimating the program's effects on children. The challenge facing the study staff was to find a way of evaluating Title I that was equally sensitive to the program's accomplishments and failures.

3. *How to overcome Congress's distrust of researchers.* At the beginning of the study, senior Congressional staff members were openly skeptical about the utility of any further research on Title I. They believed that most researchers do not take legislators' information needs seriously, and do not understand the policy process well enough to produce relevant information. This meant that, above all, the staff had to gain the professional respect of the members and staffs of the education committees at the outset.

4. *How to resist pressures from the contending parties in policy disputes.* The study mandate was originally formulated by Congressional staff members, whose principals disagreed about the merits of the Quie bill. Thus, the study's chief Congressional supporters were also strongly committed to antagonistic positions about one of the most important questions NIE had to investigate. In addition, interest groups hoped that the study results would support their positions. The study staff, under political pressure from all sides, had to gain a reputation for complete impartiality if its work were to be accepted.

5. *How to make the reports of research results useful to Congress.* Congress's dissatisfaction with prior evaluations arose partly from the reports' poor timing for legislative use, and their inaccessible language and format. Congress therefore seldom used research reports except as advocacy tools in the hands of particular members. The authors of the Title I study legislation intended the results to be a general resource to Congress, and the NIE staff knew that its work would not be considered a success unless Congress could use it that way.

The next part discusses these problems in more detail and the way NIE tried to cope with them. Some of the problems, such as advocacy pressures, are inevitable for a major program evaluation; they can be kept at bay but never eliminated.

Others may be solvable in theory, but are too complex for anyone to solve on the first try.

THE PROBLEMS AND HOW NIE DEALT WITH THEM

How to Build a Strategy of Research in Response to the Mandate

We began trying to build the research strategy months before the mandate became law, and continued to work on it for nearly two years, until long after the first research projects started. We tried to start with the obvious research projects, and the ones that would take longest to complete, but we continued trying to understand the mandate's implications as long as we had time and money available for research.

The effort had two main features. The first was to consult with researchers and interest group representatives, soliciting their advice about topics to be studied and avoided. We conducted these consultations early, and they helped shape the broad outlines of our research strategy. The second feature was continual consultation with Congressional staff in order to understand their priorities, identify information needs that were not clearly stated in the mandate, and obtain their support for projects that we wanted to include.

Consultation with Researchers and Interest Groups. During the summer and early fall of 1974, when we were writing first drafts of the research plan, some of our staff held formal meetings with researchers, representatives of state and local practitioners, and minority group leaders. When full drafts of the research plan were available, we met again with selected interest group representatives to hear their criticisms and suggestions.

Researchers mostly offered general advice, but did not suggest specific projects. Many of them had become discouraged from their own efforts to quantify the relationships between education program inputs and student achievement. They urged us to ask simpler, more fundamental questions, and to avoid sophisticated causal modeling until we understood more about program operations. Researchers associated with state departments of education and local school districts were more concrete, urging us to conduct simple descriptive studies and to employ familiar and proven research methods. They recommended that we set modest goals, especially for studies of the effects of Title I services on student achievement: These studies should acknowledge the diversity of instructional programs, and judge the effectiveness of instruction only when its characteristics are well understood. The study should pay attention to the problems of day-to-day program implementation, and try to understand the effects of federal and state regulations on local education practice.

Interest groups were concerned with protecting the future of the Title I program and their own stake in it. Their most common theme was a demand that they be consulted throughout the study. Minority groups also asked that their members be well represented on the NIE staff and among research contractors. There was no distinctive practitioner or minority group position on the problems to be studied.

The consultations with researchers and interest groups did not help much in defining research problems. Our discussions with Congressional staff, however, defined much of the research agenda.

Consultations with Congressional Staff. We began our consultations with Congressional staff by trying to make sure that we understood the language of the mandate. We knew that we had to investigate Representative Quie's proposal, to study different ways of allocating funds, and to study the effectiveness of individualized instruction. But our early discussions with authors of the statute revealed that the mandate did not include many of the questions most important to them. Congressional staffers wanted NIE to think beyond the written mandate, and to present them with a richer menu than either the mandate or their direct questions encompassed. They expected us to use creatively the great sums they had given us, to respond to their concerns and to a broader range of questions that we would develop together.

We therefore decided to take our planning orientation from discussions with Congressional staff. In our discussions, we first reached agreement about how to respond to the requirements expressly contained in the mandate; second, we asked them to identify important topics that they thought the mandate had left out; and third, we did our own analyses of Congress's information needs and proposed additional research.

Responding to the Express Requirements of the Mandate. From the statute it was apparent that Congress wanted the study to provide a clear account of how Title I—including the statute itself, the actions of federal, state, and local administrators, and the services finally delivered to children—was then operating; and to help them anticipate how changes in the statute would affect funding, administrative arrangements, and services.

Description of a program is a necessary first step toward understanding it, but entails far more than compiling a mountain of facts. Programs like Title I involve at least two, and usually three, levels of government. The federal government has many bureaucratic compartments, negotiations among which can profoundly affect the way a Congressional decision is implemented. The states and localities also have internal divisions with different functions, goals, and points of view. All of these facts affect the program's operations, and no single research project can encompass them all, nor could Congress ever hope to assimilate them. But researchers can try to explain at least the broad outlines of the program and the generic relationships among Congressional actions, management decisions by the federal bureaucracy, and the operation of the program by state and local governments.

We resolved to do that, but first we had to find a simple, intuitively satisfying conceptual framework. We saw two basic choices. One was to construct a framework based on the different levels of government that have administrative responsibilities. That approach would have led to a study that focused separately on federal, state, and local activities. The other possible framework would cut across the levels of government, through the chain of administrative actions by which the program is implemented. It would lead to a study organized around the sequential functions of funds allocation, administrative decisionmaking, and delivery of services to children. We chose the latter option because we thought a dynamic conception of the program would make it easier to anticipate the likely effects of changes in the statute. Consequently, we planned and managed the whole study around four broad topics:

- Allocation and distribution of funds;
- Relationships of federal, state, and local education agencies (school districts) in regulating and managing program activities;
- Delivery of services by school districts; and
- Changes in abilities and performance of participating students.

This framework of four topics became the organizing principle for the original research plan. When members of the House and Senate Education Committees accepted it, we decided also to build the research staff and our reports to Congress around it. The framework affected our work profoundly. It may have been one of our most important products, because it emphasized the fact that Title I is a program with many features, all of which had to be understood. It treated changes in student achievement as an important but not dominant topic. By regarding student achievement as only one element of the program, we hoped to play it down as an evaluation criterion and to make the point that a program's effects on student achievement can derive from funding and administrative arrangements as well as from the quality of instruction.

Working with Congressional Staff to Identify Important Topics Not Included in the Written Mandate. In our early negotiations with Congressional staff members, they made it clear that they regarded the written mandate as merely an illustrative list of possible research topics. Our discussions about the implications of the mandate cemented our relations with Congressional staff; they also gave us a real opportunity for leadership. Congressional staff members asked us to conduct several studies that were not expressly mentioned in the statute. Their requests, however, were very general: They identified general topics to be investigated, but left it to us to formulate research problems and propose particular projects. We took care to formulate reasonable research problems, and to present Congressional staff members with options that both bore on the topics they had identified and would lead to interesting research.

One example concerns the research on subcounty allocation of Title I funds that was reported in Chap. 6 of the NIE report, *The Title I Funds Allocation System.* That study began with a request from a senior House committee staff member for "a close look at subcounty allocation." The federal funding formula did not control the allocation of funds within counties, and regulations were very loose; the staff member had the impression that the process in some places was based on negotiation, not standard data, and might be working to the disadvantage of central city districts. He wanted to know more about the topic, but was not prepared to formulate a specific question about it. An NIE staff member was assigned to read up on subcounty allocation. He concluded that the regulations were indeed very loose, but the various allocation methods used were not necessarily creating very different patterns of funds allocation. We therefore proposed a study that would describe the range of subcounty allocation procedures then in use, describe the effects on actual allocation, and analyze possible ways of standardizing the process. Congressional staff accepted the proposal. Though all of this took place after Congress had accepted NIE's formal research plan, the new project was carried out as part of the mandated study. Several of our other projects were started in the same way.

Proposing Additional Research that We Thought Would Be Useful to Congress. Most of our consultation with Congressional staff was about research projects identified through NIE's own analysis of Congress's likely information needs. The projects were presented to Congressional staff, to find out whether they would be useful to Congress and to elicit a clear statement of interest so that they could thereafter be regarded as "mandated." Once a study was so "mandated," we could use study funds for it, and could cite Congressional interest in requesting cooperation from federal, state, and local agencies. That process continued throughout the study. The formal research plan submitted in December 1975 contained many such proposals, but others were defined thereafter.

Two papers written for us by Jim Harvey, a former Congressional staff member who had been a key participant in the 1974 House-Senate Conference negotiations on ESEA, made major contributions to that process. The first paper reviewed the legislative history of the study mandate, identified the members of Congress who expressed interest in the study, and analyzed the sources of support and opposition for the study. That paper both identified topics that were "musts" for the study, and warned about areas of Congressional conflict in which the study should operate very carefully.

The second paper discussed the political environment of Title I. It identified the long-term political commitments of key members of Congress and discussed the customary positions of important educational, minority, and public interest pressure groups. After drawing a rough "map" of such political pressures and commitments, the paper drew inferences about the political possibility of changes in the program.

The two papers identified some things that had to be done or avoided, but they could not provide a general framework for thinking about Congress's information needs. For that, we tried to think systematically about the kinds of choices that Congress as an institution is able to make. We found that it was not enough to say that we wanted to focus on "policy relevant" research. Congress's range of policymaking is narrower than the whole set of questions that could be asked about the operation of Title I. Congress can make policy in some areas (the appropriation of federal funds for example) where others have no standing, but others may make policy in some areas (such as the selection of teachers and day-to-day delivery of instruction) that are beyond Congress's control.

To understand Congress's information needs, we identified the ways that Congress can affect the operation of Title I; we then used the list of Congress's policy "tools" to identify possible research questions. Congress performs the following functions:

- Controlling the appropriations of federal funds among programs;
- Establishing formulas for the distribution of program funds among jurisdictions;
- Specifying the broad classes of purposes for which funds may be spent;
- Reviewing the regulations under which programs are administered;
- Determining the gross amounts to be spent on program administration; and
- Reviewing federal administrative performance.

Those tools are powerful, but very crude. Together they might enable Congress to exert a fine day-to-day control over events, but only with far more sustained attention to each program than individual members, or Congress as an institution, are able to give.

The simple exercise of listing Congress's policy tools made it clear that we should emphasize questions about the effects of funding formulas, regulations, and administrative arrangements, rather than questions beyond Congress's control, such as the effects of alternative teaching methods. We could not adhere to these limits rigorously, because neither we nor Congress fully understood them. Further, Congress itself had written some questions into the formal mandate about such topics as the effectiveness of individualized instruction that could bear only indirectly on any concrete decisions about the statute, regulations, or funding. Still, the focus on Congress's policy tools strongly affected our selection of research problems. For example, we did not have to include difficult—and politically problematic—basic research on the cognitive processes of disadvantaged children, and our attention was directed toward problems within the Title I program. Descriptive information, to provide Congress with a coherent picture of how the program affected the actions of federal, state, and local education agencies, was indispensable. Consequently, most of our research was primarily descriptive, especially that reported in *Compensatory Education Services, The Administration of Compensatory Education*, and *Title I Funds Allocation: The Current Formula.*

Congress also wanted information about the possibility of some narrowly circumscribed changes in the Title I program, particularly about the costs and feasibility of Mr. Quie's proposal to base funding on student achievement measures, but it did not want an exhaustive review of possible funding systems. Similarly, Congress wanted to know whether Title I regulations and management could be adjusted or strengthened with marginal commitments of new resources, but it did not want to consider a major change in the respective roles of the federal, state, and local governments.

In general, Congress was content to know about the likely effects of making marginal changes in its use of familiar policy instruments, and did not need speculative research on more fundamental or unlikely changes. The goal of most of NIE's research was therefore to understand the Title I program—how it worked, where changes were technically possible, and what their effects might be.

It is impossible to overestimate the importance of our good luck in being able to work with experienced and interested Congressional staff. The principal staff members of the House and Senate Education Committees (Jack Jennings and Chris Cross in the House, Jean Frolicher and Greg Fusco in the Senate) knew they needed more information about Title I, and were willng to work with us to get it. In the beginning, most of them were openly skeptical about whether our demands on their time could be worthwhile. We learned quickly that we should contact Congressional staff only when we had a sharply defined question or piece of information, and that they were always on the alert for signs of political bias or obscurantism on our part. We tried not to contact any Congressional staff member more often than twice a month, and prepared our agendas carefully. After enduring several months of testing—and a few rebuffs—we gained relatively easy access to key staff members. Had they been less willing to talk with us, discuss the plans, and respond to drafts and proposals, our study would not have been responsive.

The necessary discussions could not have taken place in the formal atmosphere of hearings. In such hearings, Congressional staff have time only to probe for weaknesses or make charges. Researchers need to be on their guard, and have reason to give as little information and attract as little attention as possible. Our Congressional clients never let such an adversary relationship develop.

Our experience demonstrates that researchers can do a great deal to turn difficult mandates into researchable problems, but there clearly are some impossible situations. In Chap. 5 of this report, Daniel Weiler and Marian Stearns discuss the problems that arise when legislators make unrealistic demands or provide inadequate resources. In those situations, researchers may feel compelled—by the desire to please policymakers or by competition for the opportunity to study an interesting problem—to make promises they cannot keep. With such promises, researchers can buy a brief honeymoon for themselves and their clients, but the aftermath is sure to be bitter. We were fortunate to have both a feasible mandate and a reasonable set of clients.

Ensuring that the Research Made a Fair Assessment of the Strengths and Weaknesses of Title I. Though most of Congress's questions concerned the operation of Title I and the effects of possible changes in the program, the mandate also included the term "evaluation." It instructed us to examine the "fundamental purposes of [compensatory education] programs and the effectiveness of such programs in attaining such purposes." From our early discussions with Congressional staff, it was clear that the term "fundamental purposes" was not intended to force us to produce a definitive statement about the value of Title I. But we could not pretend that the mandate was free of references to evaluation. We also knew that the original writers of the mandate could not hope to control the way that the study results were used. Because the debate about whether compensatory education had failed was then at a peak, we knew that disputants on both sides would use our results for ammunition. Whether we or the mandate writers liked it or not, our results would be seen as an evaluation of Title I.

We decided to take the "fundamental purposes" language seriously and to use it as the core of our evaluation strategy. In search of "fundamental purposes," we traced the legislative history of the program back to Adam Clayton Powell's early speeches and committee reports. Those sources confirmed that the authors' conceptions of the purposes of Title I were extremely diverse. The program was meant to establish the principle of federal aid to education, help out needy school districts, provide a symbol of federal concern for the poor, provide employment in low-income areas, revolutionize educational practice, and provide children with everything from clothing and pocket money to better instructional services. Title I, like most other major federal programs, was established through a complex legislative process that required the formation of broad coalitions of supporters. The legislative history shows that support for the program came from diverse sources. Congress was not united behind any single version of the program's goals.

The diversity of legislative purposes made it clear to us that the existing research models for evaluating federal education programs were inappropriate. Prior evaluations conducted by people who wanted to be objective about compensatory education had tried to measure the effectiveness of the program in terms of its effects on participating students' achievement test scores. As Berryman and Glennan explain in Chap. 2, earlier evaluators assumed that the Title I program was a single intervention, like a treatment in a psychological experiment, whose effects could be summarized by a simple measurement. That assumption made it possible to use readily available standard methods to evaluate the Title I program, but it did not correspond to the program's purposes. The program's goals are not limited to instructional outcomes. An evaluation that draws unfavorable conclusions about the program on the basis of student achievement outcomes ignores objectives that the Congressional supporters of Title I had in mind.

We thus had a clear mandate to evaluate the program, but no model of how to conduct such an evaluation. We tried to use our early discussions with Congressional staff to define Congress's expectations and to try out different possible responses to the evaluation requirement. It soon became clear that we did not have to produce a single "bottom line" measure of the worth of Title I. Our results might be used in evaluating the program, but Congress had no intention of delegating final judgment to us.

Perhaps our most important discovery was that Congress, not researchers, evaluates federal programs. Some of our research results might prove germane and useful in its debates, but no piece of research was likely to be conclusive. The experience with evaluations of Title I bears this out. From 1965 to the present, most researchers have drawn dire conclusions about the effectiveness of Title I, yet Congress has continued to reauthorize and expand it. Much the same is true for programs in bilingual education and Follow Through.

Congressional reluctance to act directly on the results of evaluations can be taken to mean that politicians are not wise or brave enough to act on the results of rational analysis—a conclusion that is satisfying to researchers whose findings have been ignored. Unfortunately, the opposite is far more plausible: Many research results are inadequate for evaluating federal education programs as then understood by Congress.

Because our purpose was to help Congress do its own evaluation, we had to identify Congress's main objectives for Title I. We had to present information about how closely Title I was meeting each of the objectives Congress had set for it. Naturally, everyone hoped the program would enhance student achievement, but that was not necessarily the one *fundamental* objective against which the program should be judged. Members of Congress, individually and together, had to decide what weight to give each of the objectives in judging the program.

Our solution to the problem of providing information that Congress could use in evaluating the program was summarized in the 1976 Interim Report of the NIE study:

> To identify the fundamental purpose of compensatory education, NIE studied the provisions of Title I and its various amendments, accompanying House and Senate reports, and Congressional debates. Those sources indicated that Title I of the Elementary and Secondary Education Act had three fundamental purposes:
>
> - To provide financial assistance to school districts in relation to their numbers of low-income children and, within those school districts, to schools with the greatest numbers of low-income students.
> - To fund special services for low-achieving children in the poorest schools.
> - To contribute to the cognitive, emotional, social, or physical development of participating students.

NIE's strategy for assessing compensatory education programs begins with the recognition that the program has several purposes . . . these fundamental purposes of Title I are consistent with one another, but each is not equally important to all Members of Congress. Congressional debates, and even the language of different parts of committee and conference reports, suggest that members of Congress differ over the relative importance of the respective purposes. Although some Congressional statements imply that the purposes form a hierarchy in which Title I delivers funds and services only to increase children's academic achievement (thus making the third fundamental purpose the most important), other statements make it clear that the allocation of funds and delivery of services are important ends in themselves.

The early national evaluations of Title I considered only the third fundamental purpose—contributing to children's development—and often judged program efficacy without accounting for the diverse ways in which school districts had implemented it. Those evaluations overlooked some important truths about Title I: It has several objectives, and under it school districts deliver a range of services with a variety of aims and emphases to a diverse set of beneficiaries. In contrast to earlier evaluations, therefore, NIE's strategy is designed to (1) provide clear information about what Title I is accomplishing toward achievement of each fundamental purpose, and (2) examine the implications of alternative ways of organizing the efforts of the federal, state, and local governments to achieve these purposes.

We took care to re-state the "three fundamental purposes" concept at every opportunity. By the time Congressional hearings started in the fall of 1977, it was familiar to the key staff members in the House and Senate. Congress apparently understood and accepted this approach. Despite our apprehensions, there was no further pressure on NIE to return to student achievement as the sole criterion.

Overcoming Congressional Distrust

Congressional staff members who wrote the NIE study mandate apparently did so with ambivalence. They needed research information to improve their future decisions and to help resolve political conflicts, but they had grave doubts about whether the mandate they had written would lead to anything worthwhile. Some of their misgivings concerned social scientists' ability to produce clear, comprehensible conclusions (discussed below under "Producing Usable Results"). Many of their doubts, however, stemmed from a real distrust of the motives of social scientists. They apparently believed that researchers do not take politicians seriously, and are secretly activists with their own axes to grind.

The impression that researchers do not take politicians seriously is rooted in an accurate perception of the attitudes that many researchers adopt in dealing with politicians. As we discovered in trying to adapt to the demands of the Congressional mandate, our research training strongly disposed us to consider practical politics an essentially nonintellectual activity. It was natural to draw a sharp distinction between "rational analysis" and the decision process of politicians. The distinction implies that politicians do not care about the relationship of means and ends, and are either uninterested in or incapable of fine logical analysis. As we learned in our contacts with senior Congressional staff, that implication may be correct for some politicians, but it is probably no more true of politicians in general than of professional researchers or any other group.

Many members of the NIE staff were trained as political scientists. We found it useful to think back to our early training, which taught that politicians, especially good ones, are intensely conscious of the connections between means and ends, and of the need to accomplish objectives without wasting financial or political resources. They are seldom free, however, to pursue one end to the disregard of all others, or to denominate all costs according to a simple uniform measure such as dollars of public expenditure. The essential role of a politician is to reach accommodations among people who are seeking contrary ends. To enlist support for a program with a goal he or she wishes to achieve, a politician often has to incorporate additional goals that other politicians want. The problem is to construct a program that maximizes the attainment of two or more goals, within the constraint that both (or all) of them must be served by the same instrument.[3] This requires rational analysis of a very high order, but to a naive observer the results can look thoroughly irrational. Researchers' intellectual contempt for politicians is often based on the *researchers'* failure to understand how programs are created. Congressional staff members are aware of that contempt, which plays no small part in Congress's distrust of researchers.

In dealing with Congressional staff, we genuinely believed that serving the decisionmaking process was a worthy professional activity. Some, including many other researchers, found that hard to believe. The present author once discussed the study with the May 12 Group, an informal group of distinguished educational researchers. Some members of the audience simply could not accept that description of our orientation. One member opined that our ostensible professional commitment to serving the decisionmaking process was a way of maintaining credibility while we built a case for our own policy preferences. He noted that one way to buy time for grinding an axe is to keep the axe out of sight as long as possible.

Another listener observed that the author, if sincere, was failing to do the researcher's most fundamental job, that of challenging the assumptions made by legislators and other politicians. He said that serving the decisionmaking process by providing the information most likely to be used by Congress in its deliberations limits the scope of research: It rules out fundamental questions, such as whether any level of government should have the power over people's minds that comes with controlling education.

The author had to agree with him, and observed that his attitude validated the very distrust we were trying to overcome. It is true that providing information for a specific decisionmaking process limits a researcher's scope for revolutionary thinking. It makes him a collaborator with the regime: Providing information to facilitate its decisionmaking processes implies an acceptance of its right to act in a particular area. It is important to remember, however, that no one is forced to undertake such a study, and researchers can decline to do so if they disagree with the regime's premise. In the case of compensatory education, the NIE staff accepted the legitimacy of the basic program and therefore agreed that it was worth helping Congress to define it. We did not feel morally obliged to be gadflies or adversaries to the state. Because the Congressional staff members we were dealing with disagreed among themselves about how particular decisions should be made, we hoped to avoid any pressures to produce biased reports. But we were indeed content to point our work toward the coming reauthorization of Title I and to provide evidence that others could use in making policy decisions.

---

[3] One clever way that politicians have discovered for meeting this challenge is to delegate the engineering of programs to bureaucracies. ESEA Title I is a good example. Once Congress had agreed to spend federal money on aid to education, program design was left to the Office of Education. USOE thus had to execute the compromises and live with the internal tensions that had been written into the statute. The fact that OE bureaucrats have often looked bungling and confused in the administration of Title I may very well testify that they have faithfully implemented the political accommodations reached by Congress in 1965.

Neither the author nor the critics at the May 12 Group meeting recognized a third possibility: that research problems could be identified from the current academic literature, without reference to the information needs of policymakers. Adopting that recourse would have relieved us of any choice between opposing or supporting the existing regime; if there were any political biases in our work, we would at least be spared from knowing about them. Political advocates might find ways of using our results, but as far as we would have been concerned that would be accidental.

We did not follow this third course because we were excited about the professional challenge of finding ways to make research useful in the coming Title I debates. Had the mandate been hopelessly ambiguous, internally contradictory, or technically infeasible, we would have turned to the professional literature for guidance in defining the research problems. But we had accepted an explicit and technically feasible contract from Congress, and saw no reason to violate it.

After a while, Congressional staff did come to believe that we were professionally committed to a detached service of the decisionmaking process, which made us more credible to Congress, as well as to the federal and state managers of the Title I program.

Managing Advocacy Pressure. Studies of major federal programs take place in an atmosphere of advocacy pressure. By undertaking a study of a major federal program, the researcher enters an arena where important things are at stake. Other participants will strive to protect their interests; they will be alert for any signs of hostile intent on the researcher's part and will do all they can to ensure that the research results help their own causes. We quickly found ourselves subject to advocacy pressures from Congress, interest groups, and the Executive Branch.

Pressures from Congress. Advocacy pressures from Congress were acute because the study originated in a Congressional conflict over Representative Quie's proposal to allocate Title I funds according to student achievement measures instead of according to measures of poverty then in use.

Though the mandate had several parts and a variety of Congressional supporters, its one indispensable purpose was to guarantee that the Quie proposal would be thoroughly investigated. That fact exposed the study to political conflict in two ways: First, Quie's opponents begrudged the publicity his proposal would receive by being exhaustively studied and reported, and therefore pressed NIE to hold its work on the Quie proposal to a minimum. Second, both Quie and his opponents wanted to guarantee that NIE's findings would support their positions. The mandate required

> An exploration of alternative methods, including the use of procedure to assess educational disadvantage, for distributing funds . . . to States, to State educational agencies, and to local educational agencies in a . . . manner which will . . . insure that such funds reach the areas of greatest current need . . . (P.L. 93-380, Sec. 821(a)(2)); and
>
> An analysis of means to identify accurately the children who have the greatest need for such programs . . . (P.L. 93-380, Sec. 821(a)(2)).

That placed us in the center of the conflict, where we had to manage, rather than evade, the advocacy pressures on us. We did that in three ways. The first and most important was to make it clear that research could not settle a conflict that was fundamentally political. We did so by forthrightly acknowledging that our research could not prove that low-achieving children, regardless of their income, deserved or needed special services more than low-income children. NIE's 1976 Interim Report to Congress echoed the message that we delivered in every informal conversation on the issue:

> The choice between allocation using achievement scores and allocation using poverty counts cannot be made on the basis of research results alone. It depends ultimately on a political choice about the characteristics of places and persons who are to benefit from funds the program provides. NIE's research can illuminate the practical consequences of a change in methods of funds allocation, but it cannot determine which method is "best" in a philosophical or ethical sense (p. II-14).

The second method was to make advocates on each side of the issue aware of the pressures we were receiving from the other side. Those pressures were apparent at the joint meetings we held from time to time with Congressional staff. They were underscored in private communication by vividly phrased acknowledgments of NIE's delicate position, such as "I know perfectly well that you would kill us if we biased things in Quie's favor, but so would [Quie's supporters] if we were to be biased in your way."

The third method was to avoid any hint of personal preference or bias on the part of NIE staff. We made it clear that we intended to conduct the extensive study of Quie's proposal that the mandate expressly required—but that in doing so we were fulfilling a professional obligation, and not advocating any particular conclusion.

The message was clear to Congress. The strategy of letting both sides see the conflicting pressures on the research made it possible for NIE to operate. Those methods worked in part because we were consistent in applying them. But success ultimately relied on the professionalism of the Congressional staff members who planned to use our results. Both sides recognized that results tainted by bias would be of no use to them. They understood that they had an important stake in maintaining our independence and respectability. This realization did not exempt us from scrutiny and criticism; both sides continued to apply gentle pressure, lest we forget they were there. That pressure had a benign effect because it reinforced our objectivity.

Pressures from Interest Groups. The Quie proposal was the most important source of advocacy pressure on NIE, but not the only one. A great deal of pressure came from interest groups outside Congress that had stakes in the Title I program. These included minority groups, state education agencies, teachers' unions, school board associations, and Catholics. Few of these groups had definite agendas; their main concern was to ensure that the results of our study would not weaken their positions. A few, notably parent groups and parochial school interests, wanted to make sure that we recognized the importance of their own place in the Title I program, and were content once we mounted small studies of their problems. Most simply wanted to frighten us away from saying anything critical about them.

Interest group pressures were never subtle, and generally came in the form of attacks on our honesty or professionalism. We resolved not to give in to them. We decided that appearing to be easily swayed by attacks would cause more suspicion than would mounting a stout defense. We chose the means of our defense carefully, however. We seldom engaged in direct confrontation with groups that attacked us. When under attack, our main concern was to ensure that we would not lose the support of the key Congressional staff members who were the study's main clients. In response to written attacks, we drafted very careful, balanced replies, which we delivered to Congress first and the attackers later.

The most serious attack was launched by the National Advisory Council on the Education of Disadvantaged Children in January 1975. The Council demanded that the study be halted until a new director of NIE could root out the biases in the research plan. We did not respond directly, but gave drafts of our response to Republican and Democratic committee staff members some days before we sent it to the National Advisory Council. Our concern always was to ensure that the attacks did not endanger our Congressional support. In a few cases, when the attacks were persistent or very strong, Congressional staff members would contact the attackers directly to express support for us. In one case, Mr. Quie made some remarks on the House floor that shut off an attack.

Having Congressional support did not remove the need to deal fairly with interest groups. We could not hope to maintain Congress's sympathy if we failed to take due account of interest group concerns. We therefore tried to understand and avoid aggravating the special sensitivities of the major interest groups. Our consultations during the planning phase of the study had helped us to identify "loaded" words and phrases that could elicit suspicion and opposition whenever they were used. As relative newcomers to the politics of compensatory education, we were at first unaware of the histories associated with particular terms and the freight they could carry. We soon learned, for example, that the term "district need," when used to refer to a possible criterion for allocating Title I funds, was associated in the minds of big-city education interests with an earlier New York State plan that would have reduced funding for cities. Minority group interests were also vigilant for any word or phrase that might reveal that we were operating from a "deficit model," i.e., an assumption that the failures of education are caused by the habits or abilities of minority students or their communities. We also learned that using the term "educationally disadvantaged student" could mark one as a supporter of the Quie bill, because Mr. Quie had taken care to define it to refer to students whose performance on formal achievement tests was low.

We learned to avoid such terms, and were generally able to avoid attack by inventing neutral synonyms. With one exception—the decision not to investigate aggregate "district need" indicators as possible Title I funding criteria—sensitivity over terms did not cause us to make substantive changes in the study.

These examples may appear trivial, but they underline the political sensitivity of our task. Without some degree of toleration from the interest groups, the study staff could have been forced to spend most of its time responding to attacks. In the unlikely event that the study could have been done at all in those circumstances, the reports would have been condemned by the interest groups, and would have lost much of their value as authoritative general sources of information. We therefore took great care to avoid any use of words that could incite suspicion or confer a rhetorical advantage on any of the contending sides. As one NIE staff member observed, all of us were becoming experts in propaganda through the effort to avoid rather than to create it. Taking these precautions was difficult, but it was the only way to ensure that our work could be understood for what it was rather than according to a system of political symbols that we did not understand.

None of these methods succeeded in heading off attacks from interest groups, which continued—though at a steadily declining pace—throughout the study; but we were able to keep operating.

Pressure from the Executive Branch. Throughout the study, we were able to avoid what every government program manager fears most: the imposition of review and clearance boards to diffuse the pressure on a controversial activity. Because of the short and firm deadlines for our reports, the delays caused by formal governmental or public reviews would have been disastrous. Support from the Hill was an important factor: Nobody wanted to take responsibility for delaying or derailing a study that Congress clearly wanted. Our short life span was also an asset: Since the study staff was a temporary unit with no bureaucratic future (and an untouchable budget), we were exempted from most of the pressures for review that permanent units always experience.

There was pressure on us from within the government, however. Unlike the pressures from Congress and interest groups, the pressures from the Executive Branch were not for conclusions favorable to a particular idea; they were for early release of data that could be used as private resources by officials in the Executive Branch, either for internal negotiations or to capture the initiative from Congress. We concluded the study within HEW at a time when the Department was trying to formulate its own policy about Title I. HEW wanted to improve the program and, at the same time, ensure that the Administration, rather than Congress or the interest groups, kept the initiative. Data from our studies could have been important ammunition, both for contending factions within HEW and for the Department as a whole in dealing with Congress. As a result, there was always pressure on us to provide preliminary results, do special tabulations, and contribute to internal HEW staff work. Those pressures became especially strong in the summer of 1977, when we were writing our final reports and HEW was trying to work out the Administration's proposals.

We were therefore in the middle between Congressional staff members who had initiated, negotiated, and funded the study and therefore felt that it was theirs, and HEW, especially the offices of the Assistant Secretary for Planning and Evaluation (ASPE) and the Assistant Secretary for Education (ASE), who reminded us that we were part of the Department and therefore had to act as good staff members first and contractors to Congress second. Both sides wanted the reports first, so that they could announce what they meant for policy. The relationship between the Hill and the Department was essentially competitive, and the fact that both were controlled by the same political party meant little.

Congress had protected its first access to our data by a provision in the mandate which read:

> Any provision of law, rule, or regulation to the contrary notwithstanding, such reports shall not be submitted to any review outside of the Institute before their transmittal to the Congress ... (P.L. 93-380, Sec. 811(c)).

That language controlled our behavior effectively, but it did not set well with the Department. Iris Rotberg, the deputy director of the study, had a confrontation with HEW Secretary Caspar Weinberger, who held that such an arrangement was unthinkable and therefore moot. Near the end of the study we were under constant pressure from the Department. A friend of the present author had the relevant section of the mandate photoenlarged and framed, to be hung on the wall for display to visitors who needed a reminder.

We received further help from the Hill in the form of a threat from the staff of the House Subcommittee. A member of NIE's external relations office had a copy of one of our drafts for purposes of internal review. While on a trip to the West Coast, he let a reporter take a peek at a couple of tables; they became the subject of an AP news story, and Congressional staff learned about one of our findings from the newspapers. They were mightily displeased, and threatened to rake NIE over the coals in public if it ever happened again.

We made some preliminary oral reports to HEW, taking care to talk to Congressional staff members at least a few hours beforehand. They were usually glad to get the preliminary information, but HEW staff were dissatisfied with it. They were convinced that we were sitting on mountains of useful material, holding it back until we got the reports all assembled. In one sense they were right: We had piles of computer printouts and unrelated facts that could have been used, along with some clever argumentation, to support any point an advocate wanted to make. But we were trying desperately to make a clear and defensible use of it, and we did not want to be forced by premature publicity into defending someone else's interpretation.

Much to our surprise, the part of the Department we encountered the least trouble with was the Title I program staff itself. We expected them to resist the evaluation, out of fear that it would damage Congressional support for the program or bring about criticism of their own conduct. Had they resisted, we would have found it difficult to obtain needed information about the program or to get state and local education agencies to cooperate with our data collection.

The Title I program staff dealt with us stiffly in the beginning, but it became clear that they were willing to cooperate with the study. After reading our research plan they were reassured that the study was not biased against the program. It took them longer to be convinced that we were not interested in evaluating the performance of particular administrators.

Our study of federal administration of Title I (reported in Chaps. 1-3 of *The Administration of Compensatory Education*) was the severest test of their cooperation. The Title I staff was reluctant, but willing to cooperate; other parts of the

Bureau of Elementary and Secondary Education (the OE unit that includes Title I) put up some resistance. We were able to overcome their resistance only after we had presented the plan for that project to Congress and received strong statements of interest. That made the project a "mandated" part of the study, and BESE then gave its full cooperation.

Our report writing was not a stately process. We were making major changes in the drafts until the day they were printed, and most reports went to the Hill less than ten days after we finished writing. So we were not hoarding valuable information; as soon as we knew what to make of research results, we published them.

The study staff clearly had the resources to resist pressures from the rest of the Department, and to retain control over the publication of our results. That was a unique privilege, due in part to the mandate, in part to constant pressures from Congress, and in part to the fact that our research organization was temporary, with no other projects and no bureaucratic future to protect. A permanent organization, such as the Office of Planning, Budgeting, and Evaluation (OPBE) in the Office of Education, would need to worry about its future funding and the fate of other less protected projects. Even with a similar mandate and active Congressional interest, OPBE would have found it hard to avoid responding to the Administration before it did to Congress.

While the main reports of the study were being written, the pressures within the Department made life especially difficult for the Director of NIE. Although the NIE study staff was able to rely on Congressional support to fend off attacks from the rest of the Department, the Director had to be concerned about a broader set of relations with the Commissioner of Education, the Assistant Secretary for Education, and the Secretary's office. Other NIE projects (or the Institute's budget) could be affected by their displeasure. The Director resisted those pressures and supported our determination to avoid sharing our report drafts inside the Department before they were transmitted to Congress. That required a great deal of personal courage on the Director's part; it was made only slightly easier by the fact that Congressional staff members had contacted her directly to urge that the reports be sent directly to Congress as the mandate required.

Making the Results Useful to Congress

From the beginning of the study, it was clear that Congress would consider the study a success only if the results were reported in time for use in the Title I reauthorization hearings, and in a form that Congressional staff members could readily understand. Meeting those requirements demanded intense management effort throughout the study, as well as great care in drafting the reports. It will help to distinguish the measures we took to assure punctuality from those we took to present the reports in a clear and understandable form.

Meeting the Deadlines. The mandate had established two deadlines: December 31, 1976, for an interim report and September 30, 1977, for the final products. The latter date was dictated by the reauthorization schedule for Title I.[4] Schedules for reauthorizing Title I are fixed by statute, and little can be done to relax them. If our research was to be useful for reauthorization, its results had to be available when Congress and the Administration were ready to start their review. Regardless of its quality, research that became available after decisions were made—or even after the main lines of debate had been drawn—was not going to be useful.

---

[4] A second "final" report was added later, for September 30, 1978, to permit a full report on demonstration projects that would not end until June 1978.

The question of when research is on time is not always easy to answer. A Congressional mandate can answer the question by setting a date, but even that may not cause the research to be available at the right moment. If the House had started its hearings on ESEA in the early summer of 1977, our reports would have come too late, even if we had faithfully met the statutory deadline of September 30. Happily, the Counsel of the Elementary and Secondary Education Subcommittee was interested in our reports, and got the hearings delayed until October.

For research to be used by Congress, results must be available at least a month before hearings are to begin—before witnesses are scheduled and blocks of time are assigned to particular topics. That gives the committee staff time to read the reports and decide how to use them. After receiving the NIE reports, the House decided to build the whole hearings process around them, holding approximately one week's hearings on each of the major topics we addressed. That forced everyone to pay attention to the results; it would not have been possible if our results came later. If reports become available during the hearings, they become part of a flood of information that cannot all be assimilated. After the hearings, Congress is busy with bargaining, not assimilating facts; it is then too late for research to be useful.

It is much harder to say when reports are "on time" for Executive Branch use. The Administration is able to assign staff to work on reauthorization issues for years before the hearings begin. The planning offices in HEW always want results before they can be produced: There is no one best time. From that, one may conclude that the Congressional process gives the best cues about when research should be completed.

To fit the cycle for reauthorizing Title I, the NIE study had to be completed in almost exactly three years. The mandate was enacted on August 24, 1974, and the reports to be used in Congressional hearings were due on September 30, 1977. The first six months were spent writing a research plan and making it available for Congressional review. As a result, the first Requests for Proposals (RFPs) for the study were issued in April 1975, and the first research projects started in late May of that year. They all had to be finished in about two years, by the summer of 1977, so that we could write our formal reports to Congress.

Our main response to the deadlines was to live by them. That required both political and managerial determination. On the political side, we had to resist heavy pressure from the National Advisory Council on the Education of Disadvantaged Children, which wanted us to do a several-year longitudinal study of the effects of compensatory instruction. When we pointed out that such a study would take more time than the mandate allowed, and therefore miss the reauthorization deadline, they insisted that we stand up to Congress on behalf of the imperatives of good research. We did not do that, both because we thought Congress had given us enough time to do what was necessary, and because USOE was starting the very longitudinal study (the Sustaining Effects Study) that the Council had suggested.

On the managerial side, the key to living within the deadlines was to attempt only what could be accomplished on time. That is not a profound thought, but it is an important one. It meant that we had to select our problems and methods very carefully, and we had to be sure that the interested members and staff in Congress knew the limits of what we could produce. We could not answer questions about the long-term development of children who receive compensatory services, and if we wanted to trace the development of instructional or administrative practices, we had to do it retrospectively.

The deadlines also forced us to define simple projects that could be designed, put into the field, and reported quickly. We mounted a large number of small projects, each designed to accomplish a simple objective, rather than a few complex, multipurpose studies.

That practice had several advantages. It meant that each project was simple enough for one NIE staff member, rather than a team, to monitor. Iris Rotberg and the author, as directors of the overall study, were thus able to get reliable and complete information on each study from a single staff member. Similarly, because our contractors did not need vast interdisciplinary teams of researchers, they experienced fewer managerial problems. Because projects were relatively self-contained, a problem or failure in one did not threaten the whole study. Had one project failed, we would have had a hole somewhere in our final report, but we would still have been able to give Congress most of what it needed.

We were, finally, able to conduct backup studies to protect ourselves against the possible failure of very crucial or difficult efforts. We knew, for example, that our study would be a failure if we could not produce a good report on the effects of changing the criterion for allocating Title I funds from poverty to student achievement (that is, adopting Mr. Quie's proposal). Producing such a report required some risky and difficult analysis of existing achievement test files. To protect ourselves, we mounted three different studies to produce that information. Happily, the best and most technically ambitious project—the one conducted by David Wiley and Annegret Harnischfeger of CEMREL, Inc.—produced very good results that Margot Nyitray of NIE was able to use as the basis for our report, *Using Achievement Test Scores to Allocate Title I Funds.* (We note in passing that it is politically risky for a government agency to conduct redundant research projects. Most outside reviewers point to "overlap and duplication" as evidence of bad research management. To the contrary, it is argued here that redundant research on the core problems of a mandated study can be essential to the study's success.)

Our reliance on a large number of simple studies did impose a special management burden. Each of the many small projects continually threatened to assume a life of its own. It was easy for members of the NIE research team to make decisions on particular projects without reference to the rest of the study. It was therefore necessary to Iris Rotberg and the author, as managers of the whole study, to keep the goal in view for all the members of the research team.

We accomplished that through the planning of the final report. Early in 1976, nearly two years before the main reports to Congress were due, we started outlining elements of the reports. Original drafts of those outlines were written by James Harvey, who was designated "report coordinator." These outlines included statements of the problems to be addressed, examples of the data to be provided, and identification of the projects that would provide the data. The researchers whose projects were to provide data then suggested better ways to formulate research questions and commented on whether the data required could be made available. When there were discrepancies between the report outline and the projects on which it relied, we then either revised the report outline or made changes in the relevant research projects.

By February 1977, eight months before the final reports were due, we had decided to issue the six separate reports identified in the Introduction to this essay. A senior staff member was then designated "drafter" for each of the six reports. The "drafters" produced detailed outlines, including expected table shells and figures. These were discussed in formal planning meetings, and appropriate changes were made in the outlines or research projects, or both.

The report plans changed significantly every time, and the eventual products were often quite unlike any of the plans. Report planning was an important tool for keeping the study integrated, however. There was always a working outline, which we could use as a framework for describing our work to Congress, NIE management, and the public.

From September 1976 until the main reports were due a year later, the senior NIE study staff devoted virtually full time to the preparation of reports. The mandate required an interim report to Congress at the end of December 1976. We used that report to lay out the research strategy in detail, give examples of early findings from the research on compensatory education services, and present previews of the objectives and logic of each of the six reports due the following September. Writing the interim report (*Evaluating Compensatory Education, An Interim Report on the NIE Compensatory Eduction Study, 1976*) was, for the staff, the most difficult and exhausting part of the whole study. We were forced to make explicit the heretofore loose connections between different studies, and had to settle important disagreements within the staff that we had been able to ignore previously.

The interim report also gave clear notice to Congress about what to expect. That helped Congressional staff to organize their own ideas and reduced our risk of being criticized by Congress for producing unpleasant surprises the following September.

Making the Reports Clear and Understandable. After years of trying to use the results of educational evaluations, Congressional staff members have developed an active distaste for the reporting conventions of social science. That distaste was expressed to us very early by a senior House committee staff member, who warned that Congress would ignore our reports if they were laborious or jargonized, or failed to draw conclusions. His advice to us was to be clear, brief, and definitive; not to fear arriving at a conclusion; but, of course, to be technically unassailable.

That injunction was formidable, but useful. The written mandate and our early discussions with Congressional staff members made it evident that the reports could not be simplistic. The study's sponsors had a great variety of questions, and they knew enough to distrust pat answers. They also intended to ask other researchers to comment on the technical quality of our research and the appropriateness of our conclusions. Thus, the reports could not be slick, trivial, or even very brief. We had to present a substantial body of material without either boring or overwhelming the readers. The reports had to use simple and direct language, but more important, they had to address topics of intrinsic interest to Congress and provide information that would help members of the Congress in their work.

The two most important things we did toward that end have already been discussed: focusing the study on understanding how the elements of the Title I program work together; and selecting research projects in collaboration with Congress, to ensure that the results would be germane to Congressional decisions about Title I.

These efforts pointed us in the right direction, but did not guarantee that the reports would be readable, of course. To render them so, we resolved on two further measures.

The first was to make the reports as direct and readable as possible. Clarity in this case, we decided, entailed more than avoiding arcane language and professional jargon. Above all, the writer should keep the audience in mind, and give first consideration to the audience's needs and interests, not the writer's. Accordingly, we resolved to concentrate on communicating findings and to resist the reseacher's natural inclination to recount a blow-by-blow history of the research. Policymakers need to know research results; they are less interested in the details of the research process. That may be what distinguishes a report for policymakers from one for scientists, who are interested equally in method and results. Still, because research is not likely to be useful to Congress if the scientific community condemns it, the reports must be buttressed by thoughtful and professional analysis. One solution is to relegate the professional credentials of research —compendious backup data, accounts of methodology, and the like—to appendixes or separate reports, where they are available to scientists but do not encumber the policymaker.

Our second measure was to present information in a strong interpretive framework. We assumed that policymakers are too busy for intellectual puzzles—for going over material again and again and teasing out its implications. If so, it would be a disservice to the reader if we left it up to the reader to integrate an assortment of disparate findings. Doing that for the reader, we concluded, was our most important intellectual task: finding how things fit together and explaining why data were important, thus enabling Congress to make informed decisions on changes in policy or practice. That task was far from easy. We found that it was impossible both to call attention to every finding of every project and to combine the results of all the projects into a logical whole. Selectivity therefore became necessary. We emphasized those results that bore on the questions Congress had asked us, or that we had woven into the original research strategy; we were able to accommodate new or unanticipated information only when it related to the mandate or the research strategy. That left a fair amount of data on the cutting-room floor (some of which may be reported independently by contractors or members of the NIE study staff).

In general, the need to synthesize our 35 studies put a premium on deciding what our data meant, instead of on reporting research results as if they were valuable for their own sake. Doing that made the whole exceed the sum of its parts. Important people in Congress read and understood our reports; they surely would not have read or understood the reports of 35 separate studies.

This effort was agonizing. Some of the reports went through as many as ten drafts before we were confident that the readers would know why we were presenting particular bits of information, but the effort was worthwhile. Perhaps the best thing about our reports is that they gave the staffs and senior members of the House and Senate authorizing committees a command of the issues brought before them. The House in particular used the topics of our reports as the organizing principle of its hearings on Title I in the fall of 1977. Each week of the hearings was focused on the topic of one of our reports. NIE's testimony—an oral summary of the relevant reports, with an opportunity for questions and answers—opened the hearings.[5] The Committee was able to confine subsequent witnesses to the topics of the hearings and to use the reports as checks against witnesses' statements. On occasion, Congress asked for special memoranda from NIE commenting on claims or problems raised by witnesses. Congress is seldom able to organize and discipline its flow of information in this way. Witnesses are often free to discuss whatever they like, and are able to make assertions with little fear of immediate contradiction by a generally trusted outside source.

HEW apparently had good reason to be concerned about our reporting directly to Congress. As Christopher Cross (then minority counsel of the House Education and Labor Committee) has recently written,

> By the time the Administration finally got around to formulating their position and formulating that into a bill, the results of the NIE and DECIMA studies had so thoroughly shaped the nature and content of the hearings that there were not many areas left in which the Administration could have a clear shot at shaping policy.[6]

To appreciate the institutional advantages conferred on Congress by our reports, it is important to remember that Congress received our reports hours, not days or weeks, before HEW.

Congress did not receive private advance reports, but did receive an organized body of information that could be read and assimilated before the hearings began. In the past, Congress has had to wait until the Administration gave its testimony to get the benefit of research conducted by the Executive Branch. NIE's reports may in this instance have changed the balance of power between Congress and the Executive, but they did so simply by giving Congress information in time for staff to use, and in a form that was not orchestrated to support particular proposals.

We took the precaution of avoiding making recommendations in our reports. One good reason for doing so was that few, if any, of our results led to unambiguous prescriptions. We could make "if . . . then" statements, in which the "if" was an assumption about policymakers' objectives; but we had reason to believe that policymakers were divided about objectives, and it was not in our province to make political judgments.

We had another very important reason for refraining from recommendations. We wanted to avoid the appearance of being merely another advocate or claimant. The main public role of members of Congress and their staffs is to receive demands and listen to self-interested versions of the facts. When a researcher couches his or her research findings in the form of a recommendation—which can be considered a mild form of demand—the researcher's objectivity is immediately suspect. But we could be a source of disinterested information, thereby helping members of Congress and their staffs understand and control their own world, instead of our trying to control them.

---

[5] See *Hearings before the Subcommittee on Elementary, Secondary, and Vocational Education of the Committee on Education and Labor*, U.S. House of Representatives, 95th Con., 1st sess., on H.R. 15, parts 13, 18, and 19.

[6] Christopher T. Cross, *Compensatory Education: A Congressional Perspective*, paper prepared for the National Conference on Urban Education, CEMREL, Inc., St. Louis, Missouri, July 13, 1978.

After we submitted the formal reports, Congress requested our comments on several questions. Many of the requests were for more detailed reports on findings that had caught their attention, or for analyses of problems that witnesses had identified in the reauthorization hearings. Some of the questions could be answered from data we already had. Our responses, usually in the form of letters to the Chairmen of the House and Senate Committees, were incorporated into the hearing records. Other requests required small new studies, which we conducted quickly and submitted as supplementary reports, for example, a study of the effects of school desegregation on the delivery of Title I services to children. NIE's supplementay report on that study resolved a number of disputes that had previously impeded Congressional decisionmaking about rules for the targeting of Title I funds.[7]

Congress made its most important request after the hearings were over. The Chairman and ranking minority members of the House Education Committee, Representatives Perkins and Quie, formally asked NIE to analyze the entire Title I statute in light of the study findings. They wanted a technical review of the internal logic and clarity of the statute, and other recommendations, in the form of proposed legislative language, that reflected what NIE had learned about ways to improve the program.

The request amounted to an invitation to make the kinds of recommendations that we had avoided making in the formal reports. We were heartened by the request, but it proved difficult to handle. It threatened to put the study staff, and the Institute as a whole, in an untenable position between Congress and the Administration. Because our formal reports had now been submitted, we were no longer able to decline requests for special help from the HEW Secretary's Office. On the one hand, the Secretary's staff was then conducting a department-wide effort to draft the Administration's proposals for reauthorization of Title I, and NIE was expected to participate without reservation. On the other hand, the House committee had expressly requested our recommendations as supplements to our formal reports. Like those reports, our recommendations were to be submitted to Congress without prior review in the Executive Branch. The dilemma was obvious: Our recommendations might conflict with the Secretary's, yet NIE was clearly expected to support Department policy.

Unlike most of the problems we had encountered earlier, this dilemma admitted of no direct solution. We had to find a reasonable way of serving two masters.

We first arranged to respond to Congress's request indirectly, by hiring a contractor to do the work. The contractor selected was Robert Silverstein of the Lawyer's Committee for Civil Rights Under Law. Silverstein, who had analyzed the Title I legal framework for our report, *The Administration of Title I,* by then knew more than anyone else about the inner workings of the Title I statute. His first task was to propose ways of resolving ambiguities in the law and consolidating related provisions that had previously been scattered throughout the Elementary and Secondary Education Act. He was to recommend technical improvements in the legal drafting of the statute, not substantive changes in the program. His second task was more sensitive. He was to draft amendments to the law that would correct problems in program operation identified by any of the NIE reports. Silverstein worked with NIE staff to comb the reports for possible recommendations and draft the appropriate legislative language. His third task was to draft amendments requested jointly by the Democratic and Republican staff of the House committee. The results were to be published by the Lawyer's Committee, without endorsement by NIE.

---

[7] These included "Implications of Follow-the-Child Proposals," "Indirect Costs of Administering Title I," and "Title I Funding and the Largest Cities—The Changes Since 1974."

That arrangement provided Congress with the information it needed, and it took the framing of explicit recommendations out of our hands. It was obviously not a perfect arrangement, because the Secretary's office knew about the request we had received from Congress and was aware of NIE's part in producing the Lawyer's Committee's report.

At the same time, we participated fully in the Secretary's task force. Our chief contribution was to answer questions and provide data requested by the task force. We took the initiative only to call attention to the lines of analyses that we knew would be reflected in the Lawyer's Committee report.

The arrangement appeared to work at first. There were no major conflicts between the Lawyer's Committee report and the Department's recommendations. Silverstein's redrafting of the statute was technically straightforward and uncontroversial. Soon after the reauthorization was complete, however, it became evident that the Department had been stung by NIE's working expressly for Congress. Senior officials in the Secretary's Office and USOE lobbied hard against a mandate that some members of the House had proposed for NIE to conduct a broad study of elementary and secondary school finance. As a result of HEW's lobbying, the 1978 Education Amendment assigned the mandate to the Secretary, not NIE. Congress made it evident that NIE was to have some role in the study, but left the exact arrangements to the Secretary. The intradepartmental negotiations about NIE's role proved so difficult that the whole study was delayed for several months. The study could not be done without NIE's research capabilities, but the Secretary's office was determined to prevent the development of another independent research staff. It is now apparent that a special Congressionally oriented study like ours is a foreign body in the Executive Branch and that Congress can have such studies only if it provides explicit legal and political support. Once that special relationship has expired, the organization responsible for the research is vulnerable to Executive Branch reprisals.

In the end, NIE's findings were repeatedly cited as authority for conclusions drawn in the House and Senate committee reports. The study succeeded in providing policymakers with trusted information and a common ground for discussion. It called attention to problems in the program's structure and operation, many of which were addressed in the reauthorization process. The results were not used, either by NIE or Congress, to support recommendations for fundamental changes in the program's objectives or levels of funding. They were used, however, to sharpen Congress's understanding of the Title I program and the options available for improving it.

SUMMARY AND CONCLUSION

One dual theme unites all of our responses to the five problems discussed above: Our essential strategy was to acknowledge Congress as the chief client for our work; our one irreducible goal was to give Congress the information it needed, when it was needed, and in an immediately usable form. Having Congress as a client helped us deal with each of the five problems in the following ways:

1. *Building a strategy of research in response to the mandate.* Though we put a great deal of energy into our early consultations with interest groups and researchers, the advice we got was diffuse and general. Those discussions did not provide much concrete guidance. In contrast, our discussions with Congress provided both a general conceptual framework for the study and helped us identify particular problems for research. Those discussions led us to think systematically about the political context in which Congress must work and about the tools it has for influencing policy. We were thus able to explain the study's objectives in terms that people in Congress understood and to propose studies that they thought would be helpful to them.

2. *Ensuring that the research made a fair assessment of the strengths and weaknesses of Title I.* At the beginning of the study, even though the mandate required us to evaluate the effectiveness of Title I, we had little confidence in the methods of evaluation then available in the research literature. Our early discussions with Congressional staff helped us define an alternative strategy. We learned that Congress never intended to delegate judgment on the worth of the program to a group of researchers. They knew that legislative politics had produced a diverse set of objectives for Title I, and that some members of Congress would evaluate the program on different grounds from others. That led us to define our role in evaluation in a new way: We identified the sets of objectives most often cited by Congress, and planned research to estimate the program's performance according to each. We then presented the results for each major objective, so that members of Congress could apply their own values and draw their own conclusions. We did not select among the objectives or apply our own weighting schemes. In effect, we used our research skills to supply facts that would help our Congressional clients conduct their own evaluation.

3. *Overcoming Congressional distrust.* Congressional staff members expected that we would have little respect for their needs and would try to formulate research problems to serve our own personal and professional interests. They were frankly surprised when we showed that we took their needs seriously and were not trying to advocate our own policy preferences. Adopting the client relationship with Congress was the specific antidote to Congressional distrust.

4. *Managing advocacy pressure.* The study was founded on a conflict between powerful adversaries in Congress, and its likely effect on the future of Title I made the study salient to the operators, beneficiaries, and opponents of the program. Advocacy pressures could have made the study impossible to conduct. However, it soon became obvious that both NIE and our Congressional sponsors needed the study to have a reputation for objectivity. Our Congressional sponsors were themselves in the ranks of those for and against the Quie bill, and needed us to serve as a trustworthy source of background information for their debates.

Our Congressional sponsors were also united about the desirability of improving Title I operations. Our research could help them in that effort only if it appeared fair and balanced to all of the interest groups involved. Finally, Congress had an institutional interest in guaranteeing that our results were not censored to fit the policy preferences of the Executive Branch.

Even before the study mandate was written into law, we discussed these facts with interested Congressmen and other staffs. As a result, they were careful to avoid imposing strong advocacy pressures themselves. They expected us to deal fairly with interest groups, but protected us against any partisan attacks, and they gave us the resources to resist any efforts at clearance or censorship by our superiors in the Executive Branch.

5. *Making the results understandable and useful to Congress.* Our effort to devise a responsive research strategy was an important first step toward ensuring that Congress would find our reports useful. Further, our discussions with Congressional staff helped us understand how the findings should be presented. Congress being our primary audience, our reports had to be readily understandable to the staff, written in terms Congress would find clear and familiar, and ready for use when issues were being formulated. Congress needed a synthesis of what we had learned about the program's current operation and about the effects of possible changes in it. Nothing else was germane. Methodological exposition was not helpful, nor was a litany of the blind alleys we had gone down. The reports had to be crisp and factual, and illuminate the relationships between facts. We accordingly presented our reports as syntheses of several studies, not study-by-study accounts of research procedures and data.

We also knew that members of Congress and their staffs wanted to use our results to define the important issues for future debate and evaluate the claims made by interest groups in public testimony. They did not want us to preempt their opportunity to define the issues by advancing our own prescriptions. We therefore provided facts and analyses of widely recognized alternatives, but did not present recommendations.

The value of a client relationship with policymakers is the most important lesson that we learned. Evaluators need policymaking clients, both to clarify the research problems they face, and to provide support against inevitable political pressures. In return, policymakers who are willing to act as clients increase their chances of getting useful evaluation results.

The relationship may not always be easy to establish. Ours was ready-made, because Congress had designed a far more explicit study mandate than its previous ones. Although most previous evaluations of Title I were technically based on statutory language (e.g., Section 151 of Title I, which requires the Office of Education to conduct periodic evaluations of ESEA programs), the statute contained few hints about Congress's information needs, and no clear requirements for a close working relationship with Congressional staff. Those evaluations had several audiences, including educational researchers, practitioners, federal program managers, and policymakers in HEW and Congress, but no single client. They were beset by the same technical and political problems that we faced, and had none of the advantages conferred by the client relationship.

This is not to imply that only Congress can be an appropriate governmental client for evaluation. Evaluations can certainly be done for Executive Branch policymakers, or even for program managers at the federal, state, and local levels.

However, it is difficult to imagine how an evaluation can be done well for more than one client. Because each set of policymakers has its own unique schedule for making decisions about a program, and its own policymaking tools, an evaluation tailored for one is unlikely to fit others.

Congress can be a particularly rewarding client for research, however, as much of the above discussion shows. The research results are likely to be used in decisions because Congress (in particular, the authorizing committees in the House and Senate) is the most important single institution in setting federal education policy. Congress can be a sophisticated consumer of research because key Congressional staff have a long professional commitment to education policy. Researchers can hope to maintain independence on the big issues because the Congressional clients are themselves likely to differ. Finally, members of Congress have long time horizons; if the evidence on a particular proposal is hopelessly mixed or if debate becomes acrimonious, they are often willing to defer decisions until the next reauthorization.

The Executive Branch can be a more difficult client. Its leaders are ultimately the bosses of government-employed researchers. They are thus able to censor or withhold research results (as the Secretary of HEW recently did in the case of the OE Sustaining Effects Study), and to take reprisals against researchers or agencies that do not support their own policy proposals. Because the structure of the Executive Branch is always changing, the agency that requests a study may not exist when research is completed. Even when structures are stable, personnel are not: During the life of the NIE study, there were three Secretaries of Education (and two acting assistant secretaries), three Commissioners of Education (and three acting commissioners), and three Directors of NIE. Meanwhile, none of the principal authors of the NIE mandate left the Hill. Consequently, leaders in the Executive Branch are eager to affect policy quickly, and are unlikely to wait for researchers to refine and reshape their reports. Executive Branch leaders are able to set Departmental policy and to report research results selectively in support of it.

Like the Congress, however, the Executive Branch has urgent needs for information and will still do its own evaluations, with or without special Congressional mandates. The primary client for those evaluations will be the Secretary's office (rather than the Secretary personally, whose name will not be known years in advance). When the research is planned, it should anticipate Secretarial information needs. That process should rely in part on the experience of the Department's legislative and program staffs and in part on an analysis of the issues that Congress is likely to force the Secretary to attend to.

Some evaluators may still find it impossible to establish a firm client relationship with any set of policymakers. The statutory mandate for a study may subject the research to coordination at several levels in the Executive Branch as well as require approval by Congress. In such cases, the researchers cannot focus on any one set of information needs, and cannot hope for the kind of political protection we received. For evaluators in that position, the most important lesson of the NIE study is that evaluation results are useful when they provide information relevant to a specific decisionmaking process. Since most, if not all, evaluations bear on the reauthorization of particular statutes, the decisionmaking process will be easy to identify. The nature of the reauthorization decision can imply a great deal about who will be the participants and what decisions they will have power to make. The record of past decisions on the same problem can provide information about perennial issues and topics of particular controversy.

Evaluators cannot hope to be completely insulated from politics. To conduct useful evaluations of national programs, researchers need some taste for political analysis and some respect for politicians. Although evaluation results are essentially technical, their ultimate use is political. Without some appreciation for the politics of a decision, evaluators are unlikely to serve the decisionmaking process well. They may produce irrelevant information and neglect research on topics that will be hotly debated; worse, they may fall into the hands of one side in a political dispute by unwittingly presenting results in terms that have assumed a special political coloration.

Being aware of the political context is not the same thing as being a political activist or advocate. Advocacy for either side in a debate, or for another position favored by the researcher, is professionally irresponsible, and the researcher who indulges in it is likely to reap well-deserved criticism or rejection. Political awareness—understanding the goals and strategies of the politicians who are likely to use evaluation results—is necessary and perfectly consistent with professional detachment.

Chapter 5

THE USES AND LIMITS OF EDUCATION EVALUATION AT THE STATE LEVEL[1]

by

Daniel Weiler and Marian Stearns[2]

FOREWORD

The preceding essays have provided perspectives on the evaluation of federal education programs. This essay discusses how to improve evaluation at the state level. In most states, funds used for education evaluation[3] could be spent more effectively. Neither policymakers nor evaluators are satisfied that state-level evaluations are as useful as they might be. This essay summarizes our analysis of this issue, and presents our views on three related topics:

1. What should state-funded evaluations be expected to accomplish?
2. How might the evaluation community improve the usefulness of its work?
3. What changes in state evaluation policies and priorities should be considered?

We recognize that these are complex problems open to different political and technical judgments. Nevertheless, state officials need to frame more effective policies for state-level evaluations.[4]

---

[1] This essay was originally prepared as a policy statement of the California Educational Management and Evaluation Commission, an advisory body created by the state legislature. The essay was transmitted to the State Board of Education in January 1979, and subsequently to other members of the education policy community in California. The authors were assisted by Commission consultants Peter May, Presley Pang, and Rhonda Cooperstein, who interviewed state officials and prepared early drafts. Although the essay was prepared for a California audience, we believe that other states which fund or conduct education program evaluations have similar problems; we have accordingly revised the essay to provide a more general statement on these issues.

[2] Assistant Director, Education Research Center, Stanford Research Institute.

[3] For purposes of this discussion, we define evaluation to mean applied social science research, including needs assessments, program evaluations, and some aspects of education research. Although "evaluation" is, strictly speaking, a distinguishable and more narrowly conceived enterprise, it is also the rubric under which many closely related applied research activities are perceived by decisionmakers, and is therefore used broadly here.

[4] Evaluations have important uses that are not tied directly to specific policy decisions. Such uses include issue redefinition, concept formulation, financial auditing and control, and a variety of political purposes, including program legitimation, justification or containment, and social reform. Discussion of these issues is beyond the scope of this essay, which is concerned with evaluation as an aid to decisionmaking.

CONFLICTING NEEDS AND PERSPECTIVES

States fund educational evaluations in the belief that they can provide information to decisionmakers that justifies their cost. Many evaluations compare program outcomes with program goals, to help determine whether public funds are being spent wisely. Policymakers do not rely exclusively on social science studies for information about education programs. They are aware that such studies have their limitations. Nevertheless, they are often disappointed with evaluations, and complain that their information needs are not being met. The following complaints are typical:

1. Evaluators are not sensitive to policymakers' needs. They often fail to address the right questions; promise more answers than they can deliver; produce reports late in a decision cycle; and measure program outcomes too narrowly.
2. Usually, so many caveats accompany the conclusions and recommendations of evaluation reports that their usefulness and reliability are questionable.
3. Evaluation reports are poorly written and difficult to understand. They often seem designed more to impress the evaluators' scientific colleagues than to assist decisionmakers.
4. Many evaluations appear to be of low technical quality; independent professionals often question their methods and conclusions. It is therefore not clear how much faith one should place in evaluation findings.
5. It is impossible to keep up with the flood of evaluation material; its sheer quantity limits the amount of attention that can be devoted to any one report.

Policymakers are not the only ones who are dissatisfied. Evaluators also feel that their needs are not being met, and their complaints often mirror those made by policymakers. In particular, they complain that:

1. Policymakers have unrealistic expectations. They do not understand the kinds of resources required for high-quality evaluations, or the limits on what currently available methods can accomplish. They often misuse evaluations by ignoring their caveats and limitations and reading more than is warranted into their conclusions.
2. Policymakers provide inadequate, mixed, or unrealistic signals to evaluators about their information needs. They either fail to say what kinds of information they want; ask for widely different kinds of information, making it hard to know where to focus study efforts; or overspecify data collection and analysis objectives, thereby locking evaluators into impractical study designs.
3. Policymakers tend to swamp state agency evaluators with mandated evaluations and requests for information, often to the point that agency staffs do not have adequate time or resources for detailed work on key issues.

The problems underlying these conflicting perceptions have seriously undermined the potential usefulness of many state-level education evaluations, and policymakers are often unable to benefit from studies completed at considerable cost in money, time, effort, and the good will of cooperating school districts.

We believe that these complaints are largely valid, and that policymakers and evaluators are locked into relationships which reinforce these problems. For example, policymakers may complain that they are flooded with evaluation reports, but they continue to mandate more. Evaluators, uncertain that policymakers are really paying attention, often prepare reports for the approval of their professional colleagues rather than for the policy community. In turn, this makes their work harder to understand, and increases policymakers' dissatisfaction with evaluations.

Such examples, however general, illustrate real flaws in state-level education evaluations as they are often prescribed, implemented, and used. These problems may be partly due to inherent differences between policymakers and evaluators, who have different needs and values, respond to different incentives, and use different language. Policymakers need specific information that can reduce uncertainty, couched in unequivocal language; they value action, and are rewarded by the approval of opinion leaders and the public. Evaluators need clear problem statements, and prefer the guarded language of science; they value inquiry, and are rewarded by the approval of their professional colleagues, based on accepted standards of scientific conduct.

To some extent, however, these differences can be reduced through changes in policies, practices, and attitudes. The question is: *What should policymakers and evaluators be doing differently?* We take up this question below under three headings: (1) appropriate expectations for state-funded evaluations, (2) ways in which evaluators might improve the usefulness of their work, and (3) possible changes in state evaluation policies and priorities.

APPROPRIATE EXPECTATIONS FOR EVALUATION

Policymakers must decide:

- What programs to initiate.
- What practices each program should be required to emphasize.
- How much support to allocate to each program relative to other programs with both similar and different objectives.
- How best to regulate, manage, and assist programs.

In making these decisions, what kind of assistance and advice should policymakers expect from evaluations? What should they demand of this work, and what uses should they assume it can be put to?

Although evaluations include many activities and serve many purposes, there are important limitations on their usefulness for state-level policymaking. In particular, policymakers are poorly served by evaluation information when they seek direct evidence about program results for dollars spent. We have the following views about the current utility of evaluation for each of the decision areas listed above:

Program Initiation

(Are students weak in certain subjects, implying a need for new programs emphasizing those subjects? For example, would new counseling programs for secondary schools be useful? Should new programs for the gifted be created?) Evaluations can provide rough—but useful—estimates of the quality of education on a statewide basis, and specialized needs-assessments for carefully defined target groups. Policymakers can rely on this information to help inform them about problems that new programs might ameliorate. However, such evaluations typically depend on narrow indexes of educational quality, expert opinions differ about the meaning of changes in these indexes, and these studies are helpful only as an adjunct to other sources of information generated by the political system.

Program Practices

(Should programs require parent involvement—paraprofessional aides—specially trained personnel? Should there be special program planning or local evaluation requirements? Should new school-level decisionmaking groups be created?) Decisions about what practices to require or encourage in education programs demand information about "what works" in education, and the ability to relate this information to available policy instruments. As currently practiced, evaluations cannot be counted on to provide very much of this information.[5] Although such studies can be informative and reasonably reliable in principle, they suffer from inadequate theoretical knowledge about pedagogy and organization processes, are difficult to implement in complex local settings, and are usually expensive.[6] They have also traditionally suffered from many of the problems of establishing scientifically valid inferences (see the discussion below of resource-allocation decisions), and are more likely to succeed when they are not tied directly to policy decision deadlines. As long as evaluations continue to focus mainly on gross program outcomes, information to support decisions about program practices will be scarce and unreliable.

Resource Allocation

(How much money should be allocated to Program A versus Program B? Should Program A be implemented statewide in its present form? For example, should bilingual programs be reduced, expanded, or left alone? How much should be budgeted for early childhood versus special reading programs? How much support should be provided in block grants as opposed to various categorical programs?) Policymakers who want to know whether to expand, cut back, or eliminate a program seek information on how well that program is meeting its objectives, particularly in comparison with other programs with both similar and different objectives. They therefore regularly mandate evaluations designed to assess program outcomes versus program costs. At present, however, evaluations cannot provide unambiguous findings to policymakers about the cost-effectiveness or relative value of state education programs. Informed judgments are possible, but statements about cause and effect are not.

---

[5] Studies of "what works" have traditionally been the province of education research, although this distinction is becoming increasingly blurred as national evaluations pay more attention to implementation strategies and other characteristics of program effectiveness.

[6] See Chap. 3 for a discussion of implementation problems.

Statements about cause and effect (Did program x produce outcome y?) require the use of experimental or quasi-experimental research designs to rule out competing explanations for observed outcomes. The conditions under which such designs can be implemented, and related conditions necessary for assessments of causality, are not present for studies of education programs[7]:

- Programs are not educational practices, but ways of spending money accompanied by general resource-use guidelines. Programs do not have direct effects on children; only educational practices do.
- Programs are implemented in widely varying ways at sites that have very different characteristics, and key program elements (staffing patterns, management strategies, etc.) are not standardized across sites or over time. Most differences in outcomes are a result of these site and implementation differences, which swamp program influences.
- Scientifically adequate control or comparison groups are ordinarily not available, and pre-program trends are not an adequate guide to outcomes that might have been expected in the absence of program implementation. Experimental designs require random assignment of subjects to control and treatment groups, and, for both groups, the elimination of all other influences (e.g., other programs) that could potentially affect relevant outcomes. Quasi-experimental approaches require the use of (nonrandomly selected) comparison groups, with the same treatment exposure guidelines. Because evaluations in complex natural settings cannot meet either of these requirements, there is no way to be sure of the extent to which a program is responsible for specific outcomes.
- Criteria of program success are difficult to define and use, and the quality of evaluation instruments for measuring outcomes is largely inadequate. To escape an overly narrow focus on cognitive achievement, program guidelines sometimes specify broad goals such as "responsiveness," "control," or "satisfaction." However, these goals are usually too broad to translate into precise evaluation criteria.

These problems do not mean that program evaluations are impossible, or that careful studies are useless. Experimental or quasi-experimental designs may be inappropriate for assessing state programs, but they provide a useful paradigm for standards of scientific inquiry. The test of scientific merit is not restricted to the ability to make causal or predictive statements; a rigorous and systematic appraisal of available information and careful attention to accepted canons of evidence can yield reasonable evaluative judgments, which policymakers can use as one of many sources of information that inform resource-allocation decisions. Although they may be less than ideal, such studies will be far better than no information at all, *providing* that the bases for study conclusions, and their limitations, are clearly explained.

---

[7] This is a partial list; see Chap. 2 for a more complete discussion of this topic.

Program Management

Policymakers seek information that will help them to develop program regulations, planning guidelines, assistance strategies, and criteria for grant-awards. (What eligibility rules should govern entry to learning disability programs? What kinds of plans should local parent groups develop under compensatory education programs? Should assistance resources be used to promote local incentives to increase in-service training, or to provide technical assistance teams from the state department of education? What should the rules be for deciding which schools should receive new or additional funding under school reform programs?) Such evaluations are most nearly like those required for decisions about what program practices to require, and similar comments apply. In one important respect, however, these evaluations can be easier to implement, less expensive, and more reliable than other studies of "what works" in education. Because they are most often used to support management decisions about specific program features, they can usually be exempted from the requirement, for scientifically valid inferences, that their findings be replicable under diverse conditions. Other requirements for establishing scientific validity can also be relaxed somewhat without impairing the usefulness of such work for management decisionmaking. So-called "formative evaluations"—limited findings on discrete aspects of program effectiveness—fall into this category. It should be emphasized, however, that while such work may be useful in support of program management, it is not a reliable source of information for broad decisions about program practices or resource allocation, and policymakers should not treat it that way.

In sum, policymakers should not expect evaluations to provide unambiguous answers about the effectiveness of a state's educational programs. Moreover, as long as most evaluation resources are committed to large, formal studies of program cost-effectiveness, little useful information will emerge on what program practices to emphasize. Evaluations today are most useful in helping to identify problems that might require new program initiatives, and in supporting management decisionmaking.

In the following pages, we consider briefly how evaluators might improve the usefulness of their work despite the problems described here, and present some suggestions for state evaluation priorities that emerge from this analysis.

IMPROVING THE USEFULNESS OF EVALUATIONS[a]

Evaluators often complain that policymakers have exaggerated expectations of what evaluations can do, and impose unrealistic demands on evaluation professionals. There is substantial truth to this complaint, but much of the fault lies with the evaluation community itself, which has repeatedly oversold its capabilities and promised more than it could deliver. In fact, while some studies are manifestly more competent than others, many so-called quality problems can be traced to exaggerated expectations of what evaluations can accomplish. Evaluators could make their work more credible by providing accurate, clearly stated, suitably modest descriptions of what policymakers can expect to get from evaluations. These statements should occupy a prominent place in every evaluation proposal or prospectus. In addition, every evaluation report should include:

---

[a] This discussion is confined to general recommendations; technical advice on how to improve the quality of evaluation studies is beyond the province of this essay.

- A summary of study objectives and assumptions.
- A description of problems encountered during the study, the strengths and weaknesses of the study design, and the limitations of study techniques and data, together with an assessment of the reliability of study findings.
- An explanation of the meaning and implications of study findings, together with a summary of possible alternative explanations and an assessment of their probable validity.
- Either a policy recommendation based on the evidence developed in the report, a discussion of alternative policies and their possible consequences, or a statement that the evidence is not sufficient to support any policy advice, together with a description of what additional information is needed.

Evaluators also complain that policymakers do not clearly indicate their information needs, and it is often suggested that evaluations would be more useful if policymakers defined their objectives more precisely. On the whole, we are not sympathetic to this view. Although a better understanding by policymakers of the limits and uses of evaluation may lead to more carefully drawn evaluation mandates, it is unrealistic to expect program objectives that are formulated in the give and take of the policy process to be restated to satisfy the needs of evaluation studies. Ambiguous goal statements are necessary for legislative compromise; it is up to evaluators to translate such statements into researchable objectives. This process should take place in five stages[9]:

1. Tentative decisions by the evaluator about policy issues and research goals.
2. Evaluator discussion with agency and legislative staff to clarify and enlarge initial understandings.
3. Preparation of a research design, including a statement of the policy issues and research questions, and development of the research approach.
4. Presentation of the design to agency and legislative staff (and interested legislators) for further comment.
5. Decision on a final research design to guide the study.

These activities should help evaluators understand the issues that interest the policy community, while maintaining their operational control of the research agenda. Where the policy interests of study clients remain ambiguous despite this interaction, evaluators must nevertheless take the initiative in deciding what questions their study should address. By doing so, they may risk the criticism of policymakers whose own understanding of program goals leads to different conclusions, but failure to take this initiative exposes evaluators to the far more serious error of undertaking a study without clear goals.

Evaluators should not hesitate to study education progams with methods that depart from experimental design. The terminology of experimental method is so deeply ingrained in popular conceptions of what constitutes "sophisticated" research that policymakers have come to expect evaluations to embody some application of this method as a hallmark of professionalism. Consequently, the trappings of experimental design have become a standard feature of many evaluations, even though adherence to such designs in practice (as we have seen) is almost always inappropriate and likely to do more harm than good. Anthropological approaches, comparative case studies, process analyses, or other scientifically sound methods are preferable to misapplied experimental designs that yield debatable statistical inferences.

---

[9] See Chap. 4 for a more extensive discussion of this topic.

Finally, it is impossible to overemphasize the virtues of clear, jargon-free writing. No evaluation should be sent to policymakers without an executive summary that presents study assumptions, findings, and conclusions in a form that a lay audience can understand.

EVALUATION POLICIES AND PRIORITIES

What we have said suggests a number of changes in state evaluation policies and priorities that might help to improve the quality and usefulness of evaluation work. Because such changes should be the subject of extensive discussions between the evaluation and policymaking communities, we offer for state agencies' consideration the following tentative general recommendations:

1. Studies that use a single outcome score to judge the relative values of programs, without regard to different program goals or approaches, are of little value. Large-scale, summative evaluations should be reconceptualized as less ambitious studies. Such studies could, for example, describe how programs are being implemented in school districts, present carefully justified judgments about the relationship of programs to changes in educational treatments that may be affecting children, and assess the usefulness of programs to districts and schools.
2. In states that presently require the annual evaluation of all or most state programs, the number of mandated evaluations in any given year should be drastically reduced; many programs should be evaluated less frequently. Moreover, once a schedule of evaluation work at a specified budget level has been agreed to, no additional evaluations should be mandated unless supplemental funds are allocated to support the new work.
3. Resources conserved by reducing the scope and number of program evaluations should be transferred to studies of the effectiveness of various strategies and practices *within* programs, in support of program practice decisions. (Such work would also be valuable at the local level, and could eventually represent a visible source of state assistance to local efforts at self-improvement.)
4. Legislation and agency Requests for Proposals addressed to universities and private research firms should not specify evaluation designs. Legislative language should be confined to statements of the policy issues to be resolved, should not require or imply that experimental (or any other)

designs are necessarily expected, and should not specify evaluation data collection and analysis requirements. Legislation *should* require, however, that evaluation reports include the materials described above (study assumptions, design strengths and weaknesses, data limitations, study conclusions, etc.), as well as executive summaries.

5. To improve the usefulness of evaluations for educational policy and management, state boards of education should assess the effectiveness of their states' current evaluation practices. Advisory commissions or technical consultants could be used to review agency research goals, evaluation plans, and requests for proposals, as well as the design and execution of research both by state agencies and outside contractors.
6. There should be continuing strong support for evaluations that can assist program initiation and program management decisions; however, the use of indexes of educational quality in addition to achievement tests should be encouraged.

The limitations on what evaluations can accomplish do not necessarily lead to the conclusion that evaluation activities are largely a waste of time and money. They point instead to the importance of acquiring a more realistic view of how evaluations can be helpful to policymakers, and of reorienting state evaluation priorities in order to take better advantage of what these studies can offer.